Cognitive Engineering Laboratory

Research on the Characteristics of Long-Term Adaptation (II)

Xinyao Yu, Gerard Torenvliet, & Kim J. Vicente

CEL 98-06, Volume 2

Final Contract Report
September, 1998

Prepared for
Japan Atomic Energy Research Institute

Cognitive Engineering Laboratory • University of Toronto • Department of Mechanical & Industrial Engineering
5 King's College Rd. Toronto, Ontario, Canada M5S 3G8
Phone: (416) 978-7399 Fax: (416) 978-3453
Email: benfica@mie.utoronto.ca URL: www.mie.utoronto.ca/labs/cel/

Director: Kim J. Vicente, B.A.Sc., M.S., Ph.D.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Mechanical & Industrial Engineering, and is one of three laboratories that comprise the U of T Human Factors Research Group. CEL began in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments, with a particular emphasis on power plant control rooms. Professor Vicente's areas of expertise include advanced interface design principles, the study of expertise, and cognitive work analysis. Thus, the general mission of CEL is to conduct principled investigations of the impact of information technology on human work so as to develop research findings that are both relevant and useful to industries in which such issues arise.

Current CEL Research Topics

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Defense and Civil Institute for Environmental Medicine, Honeywell Technology Center, Japan Atomic Energy Research Institute, Natural Sciences and Engineering Research Council of Canada, Rotoflex International, and Westinghouse Science & Technology Center. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Recent CEL projects include:

- Studying the interaction between interface design and adaptation in process control systems.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing safer and more efficient interfaces for computer-based medical devices.
- Designing novel computer interfaces to display the status of aircraft engineering systems.
- Developing and evaluating advanced user interfaces (in particular, transparent UI tools) for 3-D modelling, animation and painting systems.

CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente at the address printed on the front of this technical report

ABSTRACT

This final contract report describes the findings from the second year of a two-year research program investigating the characteristics of long-term operator adaptation in nuclear power plants (NPPs). The report consists of two volumes. This document, volume 2, describes the results of a systematic literature review that was conducted to understand the limitations of well-known statistical analysis techniques, namely null hypothesis significance testing and ANOVA. This review uncovered six major points:

1. Averaging across subjects can be misleading.
2. Strong predictions are preferable to weak predictions.
3. Constructs and measures should be distinguished conceptually and empirically.
4. Reliability and magnitude should be distinguished conceptually and empirically.
5. The null hypothesis is never true.
6. One experiment is always inconclusive.

Based on these insights, a number of lesser-known statistical analysis techniques were identified to address the limitations of more traditional techniques. Several of these lesser-known techniques were applied to data from previous experiments conducted for JAERI. The results of these analyses have confirmed results obtained previously, added to our understanding of our existing dataset, and have suggested methodological refinements for future experimentation. Thus, these less-traditional analysis techniques are perhaps better suited to the purposes of cognitive engineering research than traditional statistical analysis techniques.

ACKNOWLEDGMENTS

This research project was sponsored by a contract from the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, Contract Monitor), as well as research and equipment grants from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Ian Spence and Dr. Tanabe for their contributions and help.

TABLE OF CONTENTS

Abstract i
Acknowledgments ii
Table of Contents iii
Table of Figures iv
Table of Tables v
Overview 1
Introduction 1
Literature Review 3
1. Averaging across subjects can be misleading. 4
2. Strong predictions are preferable to weak predictions. 7
3. Constructs and methods for measurement should be distinguished conceptually and empirically 9
4. Reliability and magnitude should be distinguished conceptually and empirically. 14
5. The null hypothesis is never true. 16
6. One experiment is always inconclusive. 20
Application to Previous JAERI Experiments 23
Introduction 23
Techniques Applied 23
Confidence Intervals and Graphical Analyses of Variance 24
Power Analysis 25
Bayesian Methods to Assert Largeness or Smallness 27
Results 35
Background 35
Confidence Intervals and Graphical Analyses of Variance 39
Power Analysis 42
Bayesian Methods to Assert Largeness or Smallness 54
Conclusions 63
Conclusions 64
Postscript 65
References 66

TABLE OF FIGURES

Figure 1: Learning curve averaged over six subjects..... 4
Figure 2: Learning curve for one of the six subjects. 5
Figure 3: Hypothetical example showing how confidence intervals reflect different degrees of measurement precision. 19
Figure 4: 80% CIs about effect sizes for JAERI II trial completion time, by effect. 35
Figure 5: Graphical ANOVA on interface effect for JAERI II completion time. 41
Figure 6: Graphical ANOVA on training effect for JAERI II completion time. 41
Figure 7: Effect size analysis on JAERI I INT effect. 56
Figure 8: Effect size analysis on JAERI II INT effect. 56
Figure 9: Effect size analysis on JAERI II TRAIN effect. 57
Figure 10: Effect size analysis on JAERI II mini experiment INT effect. 57
Figure 11: Effect size analysis on JAERI IIIa INT effect. 58
Figure 12: Effect size analysis on TCT across experiments. 58
Figure 13: Effect size analysis on CTV across experiments. 59
Figure 14: Effect size analysis on DET across experiments. 59
Figure 15: Effect size analysis on DA across experiments. 60
Figure 16: Effect size analysis on DGT across experiments. 60
Figure 17: Effect size analysis on CT across experiments. 61
Figure 18: ANOVA *p*-values vs. effect sizes..... 63

TABLE OF TABLES

Table 1: A multitrait-multimethod matrix 12
Table 2: A multitrait-multimethod matrix 13
Table 3: Partial ANOVA table for JAERI II diagnosis score. 32
Table 4: Integrated summary of the three-year research program 37
Table 5: ANOVA table for JAERI II trial completion time. 40
Table 6: Power analysis on TCT for JAERI I. 43
Table 7: Block definitions 44
Table 8: Power analysis on TCT by block for JAERI I. 44
Table 9: Power analysis on JAERI I CTV. 45
Table 10: Power analysis on JAERI I DET. 45
Table 11: Power analysis on JAERI I DA. 46
Table 12: Power analysis on JAERI I DGT. 46
Table 13: Power analysis on JAERI I CT. 46
Table 14: Power analysis on TCT for JAERI II. 47
Table 15: Power analysis on CTV for JAERI II. 47
Table 16: Power analysis on DET for JAERI II. 48
Table 17: Power analysis on DA for JAERI II. 48
Table 18: Power analysis on DGT for JAERI II. 48
Table 19: Power analysis on CT for JAERI II. 48
Table 20: Power analysis on TCT for JAERI II no training data. 49
Table 21: Power analysis on CTV for JAERI II no training data. 49
Table 22: Power analysis on DET for JAERI II no training data. 49
Table 23: Power analysis on DA for JAERI II no training data. 50
Table 24: Power analysis on DGT for JAERI II no training data. 50
Table 25: Power analysis on CT for JAERI II no training data. 50
Table 26: Power analysis on TCT for JAERI IIIa. 51
Table 27: Power analysis on CTV for JAERI IIIa. 51
Table 28: Power analysis on DET for JAERI IIIa. 51
Table 29: Power analysis on DA for JAERI IIIa. 51
Table 30: Power analysis on DGT for JAERI IIIa. 51

Table 31: Power analysis on CT for JAERI IIIa. 51
Table 32: Power analysis on TCT for JAERI IIIb. 52
Table 33: Power analysis on CTV for JAERI IIIb. 52
Table 34: Power analysis on DET for JAERI IIIb. 53
Table 35: Power analysis on DA for JAERI IIIb. 53
Table 36: Power analysis on DGT for JAERI IIIb. 53
Table 37: Power analysis on CT for JAERI IIIb. 53
Table 38: 1-*df* effect size analyses. 55

OVERVIEW

A predictive model of human cognitive behaviour, which includes the mental strategies used in emergency situations in nuclear power plants (NPPs), is needed in the design and evaluation of human-machine systems. To achieve this objective, a profound understanding of the characteristics of human operators' long-term adaptation to the major behaviour shaping constraints in complex systems is essential. This year's projects builds on the work conducted in last year's project for JAERI by further investigating the usefulness of novel measures of operator adaptation. The results obtained will be useful for the development of a model of human operator cognitive behaviour and of criteria for design and evaluation of human-machine systems.

The work conducted during this project is documented in two volumes. Volume 1 describes the results of analyses of data from tuning and fault trials in a 6-month longitudinal study of operator adaptation using the novel measures of adaptation that were developed in last year's contract. In addition, volume 1 also provides a theoretical integration of this year's findings and those obtained in last year's project on long-term operator adaptation. This document, volume 2, describes the results of a systematic literature review that was conducted to understand the limitations of well-known behavioural statistical analysis techniques, such as null hypothesis significance testing and analysis of variance. In addition, this document describes a number of lesser-known statistical analysis techniques that were identified to address the limitations of the more traditional techniques, and shows how these techniques were applied to data from previous experiments conducted for JAERI.

INTRODUCTION

Whether it be in experimental psychology or human factors engineering, the statistical analysis of data almost always relies on two related techniques, null-hypothesis significance testing (NHST) and analysis of variance (ANOVA). All of us have been taught these techniques and we have been told that they are the scientific way to analyse data statistically. These techniques are so commonly used and so widely accepted that we frequently apply them to our 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, p. 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 & Hunter, 1997, p. 39)

> Hypothesis testing is the wave of the past (and never should have been a wave at all). (Loftus, 1993b, p. 255)

> The common belief that the precise quantity [p <] .05 refers to anything meaningful or interesting is illusory. (Loftus, in press, p. 165)

> 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, p. 817)

> 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, p. 149)

It might be tempting to dismiss these strong criticisms as uninformed, "fringe" opinions. However, the authors that we cite include very well-known and highly-respected researchers. Thus, their criticisms should cause us to at least reflect upon, if not revise, the way in which we statistically analyse our data.

Critiques of NHST and ANOVA go back at least to the 1960s (e.g. Bakan, 1966; Lykken, 1968; Meehl, 1967; Rozeboom, 1960), have resurfaced periodically in the 1970s and 1980s (e.g. Hammond, Hamm, & Grassia, 1986; Hammond, Hamm, Grassia, & Pearson, 1987; Meehl, 1978; Rosnow & Rosenthal, 1989)], and have appeared with increasing frequency and cogency in this decade (Cohen, 1990, 1994; Hammond, 1996; Harlow, Mulaik, & Steiger, 1997; Loftus, 1991, 1993b, 1995, in press; Loftus & Masson, 1994; Loftus & McLean, 1997; Meehl, 1990). These critiques have been met with rebuttals (Chow, 1996; Hagen, 1997; Harlow et al., 1997; Serlin & Lapsley, 1985) but there is a growing consensus that there are sound reasons to justify discontent

with traditional methods of statistical data analysis. To be clear, our purpose in this report is not to make an original technical contribution to this literature nor is to dismiss the use of the traditional techniques. Instead, we aim to bring the practical implications of this literature to the attention of the human factors community so that we can suggest some alternative, complementary ways of analysing data statistically. Our experience has been that most human factors professionals are not aware of these limitations. Thus, they could benefit both from a deeper understanding of them as well as from the knowledge of a broader set of analytical techniques that address these limitations.

The remainder of this report is organised as follows. First, the results of our literature review will be presented. We will identify six major points and the implications that these points have for alternative methods of statistical data analysis. Second, we will use some of these alternative methods to analyse data from previous experiments conducted for JAERI. In addition to providing general examples of how these lesser-known techniques can be applied in practice, these analyses also reveal more specific insights that are very useful for future research to be conducted for JAERI.

LITERATURE REVIEW

The results of the literature review are organised according to six major points:

1. Averaging across subjects can be misleading.
2. Strong predictions are preferable to weak predictions.
3. Constructs and methods for measurement should be distinguished conceptually and empirically.
4. Reliability and magnitude should be distinguished conceptually and empirically.
5. The null hypothesis is never true.
6. One experiment is always inconclusive.

Although some of these points may seem self-evident, our review will show that they are frequently not heeded by psychologists and human factors engineers. 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 perhaps more valuable, set of insights into our data.

1. Averaging across subjects can be misleading.

We will begin by discussing an issue with which many researchers are familiar but that is nevertheless frequently overlooked in the statistical analysis of data. Both NHST and ANOVA involve averaging across subjects, and as a result, it is commonplace for researchers to assess statistical significance at an aggregate level of group means. Yet, the act of taking an average only makes sense if the samples being aggregated are not qualitatively different from each other. Without looking at each subject's data individually, we do not know whether 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 subject in the group.

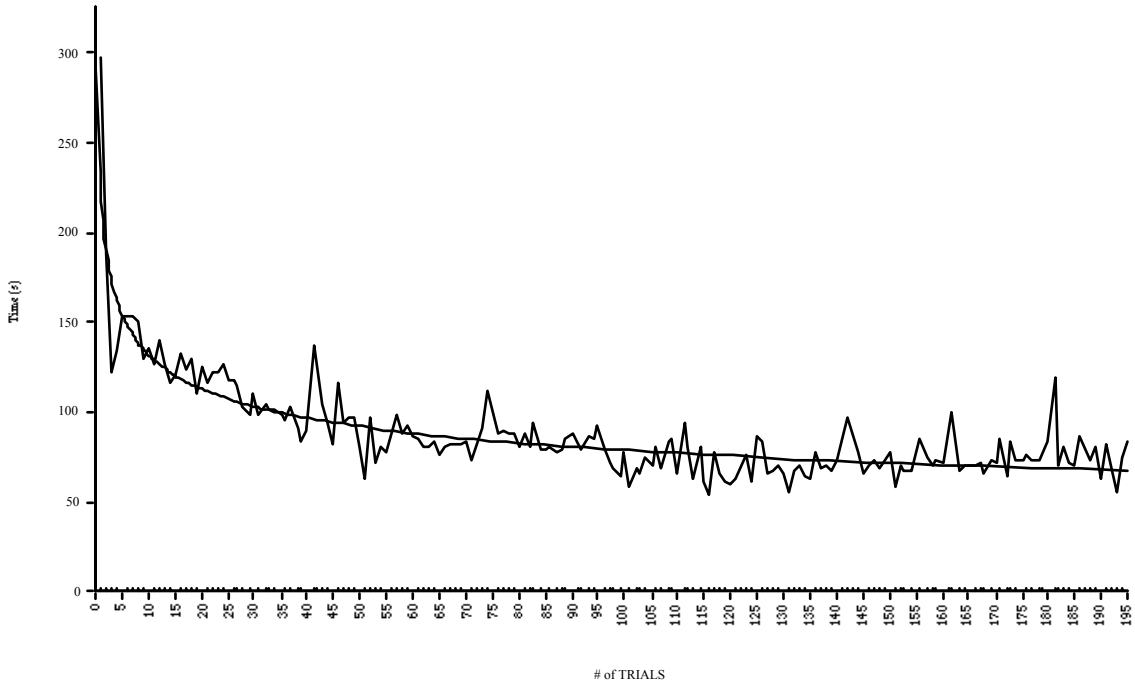

Figure 1: Learning curve averaged over six subjects (Christoffersen, Hunter, & Vicente, 1994).

Data from a longitudinal study conducted by Christoffersen, Hunter, and Vicente (1994) can be used to illustrate this point in a very demonstrable fashion. Figure 1 is 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 subjects. A power law fit has been superimposed on the aggregate data in Figure 1. Based on visual inspection alone, we can see that there is a good fit between

the data and the power law curve. A regression analysis showing a r^2 value of 0.74 confirms this impression. We might conclude from this aggregate-level analysis that these data provide support for the power law of practice (Newell & Rosenbloom, 1981). However, such a conclusion might be premature. Without looking at each subject's data, we do not know whether the elegant power curve fit would provide an equally good account of the skill acquisition behaviour of each individual.

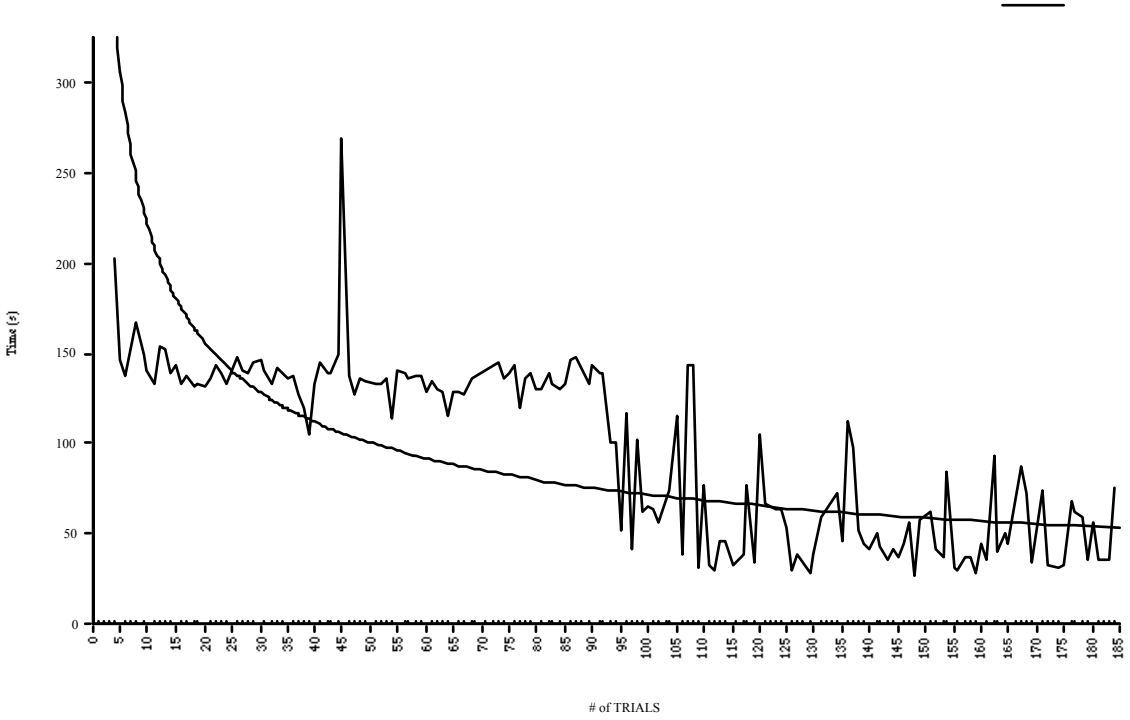

Figure 2: Learning curve for one of the six subjects (Christoffersen et al., 1994).

Figure 2 shows the learning curve data for one of the six subjects. Again, a best fit power law curve has been superimposed over the raw data. It is obvious that the degree of fit between the power law of practice and this subject's data is very poor. Thus, to use the average as a basis for generalising to individuals would be very misleading in this case.

Plateaus in learning curves and the dangers of aggregating data over subjects are hardly new insights (Bryan & Harter, 1897, 1899; Woodworth, 1938). Yet, as Venda and Venda (1995) have recently pointed out, these insights are still frequently ignored by many, although by no means all, human factors researchers. We believe that, in part, these oversights result from the fact that NHST and ANOVA induce us to aggregate our data over subjects. Thus, we must make

a special effort 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 subjects to heart can actually take us to new and perhaps more powerful ways of analysing data. In cases where we use a within-subjects design, we can use each individual subject as an experiment and see if the theoretical predictions being tested hold for each individual. An example of this type of test is provided by Vicente (1992) who compared the performance of the same group of subjects with two different interfaces, one labelled P and the other labelled P+F. There were theoretical reasons for hypothesising that the P+F interface would lead to better performance than the P. However, rather than just seeing if the means of the two conditions differed, Vicente also conducted a more detailed analysis to see if the theoretical prediction held for each and every subject. The number of subjects for whom the hypothesised relationship (P+F > P) held was counted and then this count was analysed statistically by conducting a sign test (Siegel, 1956). In one analysis, the P+F interface led to better performance than the P for 11 out of 12 experts, a result that is significant at the $p < 0.001$ level.

This example is important for at least two reasons. First, in at least some applied situations, it may be more important for us 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 we are testing the performance impact of an advanced control room for a NPP. Are we more interested in knowing whether the mean performance of the new control room is better than that with the old, or are we more interested in knowing the proportion of operators for which performance with the new control room is better? It seems that the latter would be more valuable. After all, a NHST or 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. Second, this example also shows that non-parametric tests (e.g., the sign test and the χ^2 test), which are usually thought to be weaker than parametric tests, can actually be used in innovative ways to test strong predictions. This topic is discussed in more detail next.

2. Strong predictions are preferable to weak predictions.

Empirical predictions can be ordered on a continuum from strong to weak (Vicente, in press). At the strong end, we have point predictions. To take a hypothetical example from physics, a theory might predict that the gravitational constant, G, should be 6.67×10^{-11} $N \cdot m^2/kg^2$. An experiment can then be conducted to see how well the data correspond to this point prediction. Slightly farther along the continuum, we find interval predictions. To continue with the same example, a different theory might only predict that 6×10^{-11} $N \cdot m^2/kg^2 < G < 7 \times 10^{-11}$ $N \cdot m^2/kg^2$. 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, we find ordinal predictions. For example, a third theory might only predict the direction of the force of gravity. In this case, all we would know is that gravity pulls objects towards, rather than away from, the earth. Finally, at the weak end of the continuum, we have categorical predictions. For example, a very primitive theory of physics might just 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 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 human factors research, 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., NHST and ANOVA) govern the strength of the predictions they make. And because NHST and ANOVA are 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), psychologists and human factors researchers generally restrict themselves to testing categorical predictions, the weakest area of the continuum and thus indicative of an 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 about gravity cited above shows just how weak such a test really is. 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 reasonably be argued that most areas of human factors research have not reached the level of theoretical maturity to make point or interval predictions. Even if this is so, it does not mean that we cannot be more ambitious than we have been in the past. Rather than letting our 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. The innovative work of Hammond, et al. (1987) provides an example of how we can begin to make stronger predictions and how these can be tested using untraditional statistical analysis techniques.

Hammond et al. (1987) 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 of highway engineers (i.e., intuition vs. analysis). There were three levels for the depth task characteristics dimension: a) an aesthetics task that was intended to induce intuition; b) a 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 highway engineer experienced each of the nine combinations of depth and surface task characteristics.

From a traditional perspective, this experimental design fits very neatly into a between-subjects 3 x 3 ANOVA. However, analysing the data in this fashion would only allow the experimenters to test some comparatively weak predictions. Statistically significant results would merely suggest that the effects of the independent variables (or their interaction) is unlikely to be zero. This 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

driving their research. Note that this is a much stronger prediction. Instead of just hypothesising that the effect was different from zero, Hammond et al. were committing to one specific ordering of their experimental conditions. And because there was a total of nine conditions in their experiment, there are many possible orderings that could conceivably occur (9! = 362,880). Yet, only one of these orderings is perfectly consistent with the prediction they were making. Second, instead of testing this ordinal prediction at the level of a group aggregate, they tested it at the level of each individual subject. 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. Third, to test the predicted ordering on a subject by subject 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 paper, 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 group.

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

3. Constructs and methods for measurement should be distinguished conceptually and empirically.

Even if we are able to make and evaluate stronger predictions, the level of our science is only as good as the empirical methods we use. 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 (cf. Xiao & Vicente, 1997). As Campbell and Fiske (1959) pointed out in their seminal paper almost 40 years ago, there are certain basic criteria that must be met before we can confidently interpret a pattern of experimental results in a meaningful fashion. Among the most important criteria are reliability, convergent validity, and discriminant validity. 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, we are instead observing the effects of the way in which the construct was measured — a much less interesting phenomenon, unless one is a methodologist. Finally, discriminant validity refers to the extent to which different 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, we are instead observing similar effects caused by the method of measurement.

A few hypothetical human factors examples can help make these abstract concepts more concrete. Say we are conducting an empirical investigation of the interaction between spatial ability and mental workload for a particular task context. How could the three criteria identified by Campbell and Fiske (1959) be operationalised? Beginning with the issue of reliability, whatever method we use to measure each construct should lead to consistent results under comparable conditions. For example, our 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 our experiment. Moving on to convergent validity, different methods of measuring the same construct should lead to consistent results under comparable conditions. For example, if we have two different methods for measuring mental workload (e.g., a computer-based version and a paper-based version of a subjective rating scale), we would ideally like those methods to give the same results for the same subject for a particular trial. If the two methods give different results, then the variance in our data is being caused by the method of measurement. In such a case, we cannot make any confident inferences about what we are really interested in, namely the construct of mental workload. Thus, like reliability, convergent validity is a prerequisite for sound scientific knowledge. As for the third criterion of discriminant validity, the same measurement methods should lead to different results for different constructs of interest. For example, a computer-based test of spatial ability should be highly correlated with a paper-based

test of spatial ability and not correlated at all with a computer-based assessment of mental workload. If this criterion is not met, then we have the case of too high a correlation between tests that are intended to measure entirely 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 we observe can be safely attributed to the content of the constructs in which we are interested 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 our results, we need a way to evaluate reliability, convergent validity, and discriminant validity empirically. To achieve this goal requires that any one experiment have at least two constructs and at least two methods of measurement. Campbell and Fiske proposed an analysis technique based on these insights that can allow experimenters to determine whether in fact they are measuring what they wish to measure 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 has been extended by Hammond, Hamm, and Grassia (1986) so that it can be applied to a much wider range of behavioural 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 hypothesised 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 surely true of human factors; studies using the MTMM technique are exceedingly rare. Researchers tend to analyse their data using the familiar NHST and ANOVA techniques. However, these 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 or not. Hammond et al. make a very strong case that this situation makes it exceedingly difficult to develop a cumulative scientific knowledge base. Instead, what we get is conflicting findings in any given literature 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 means of remedying this situation.

Lee's (1992) investigation of the relationship between operator trust, self confidence, and the use of automation is one of the very few applications of MTMM in the human factors literature. 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. Such a design allows us to construct the matrix shown in Table 1. 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.

Table 1: A multitrait-multimethod matrix relating trust and self confidence measured by subjective scales and frequency of monitoring behaviour for Lee's (1992) study. A + indicates that a high correlation is expected in that cell (a sign of convergent validity). An X indicates that a very low correlation is expected in that cell (a sign of divergent validity). SS is an abbreviation for 'subjective scales', and MB is an abbreviation for 'monitoring behaviour'.

		Trust		***Self-confidence***	
		SS	**MB**	**SS**	**MB**
Trust	**SS**				
	MB	+			
Self-confidence	**SS**	X	X		
	MB	X	X	+	

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. We 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. We should expect to see a high correlation value in this cell as well 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. We should expect to see very low correlation values (indicated by a X) in these cells. For example, ratings of self confidence on a subjective scale and ratings of trust on a subjective scale should not be correlated because they are measuring different constructs. If the data turn out to be correlated, then we 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 of the results 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). We would expect to see a low correlation here because different constructs should lead to different results. The fact that there is a 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. We 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.

Table 2: A multitrait-multimethod matrix relating trust and self confidence measured by subjective scales and frequency of monitoring behaviour (Lee, 1992). The values are the means of z-transformed correlation coefficients of individual operators. A + indicates that a high correlation was expected in that cell (a sign of convergent validity). An X indicates that a very low correlation was expected in that cell (a sign of divergent validity). Abbreviations are as in Table 1.

		Trust		***Self-confidence***	
		SS	**MB**	**SS**	**MB**
Trust	**SS**				
	MB	0.15 (+)			
Self-confidence	**SS**	0.42 (X)	0.04 (X)		
	MB	-0.07 (X)	-0.08 (X)	0.04 (+)	

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

4. Reliability and magnitude should be distinguished conceptually and empirically.

It is a truism in human factors engineering that statistical significance is not the same as practical significance. This truism has a sound basis in statistics. Statistical significance is a measure of reliability, and thus indicates the degree of uncertainty in our results. In contrast, effect size (Abelson, 1995; Cohen, 1988, 1990, 1994; Rosnow & Rosenthal, 1989; Rouanet, 1996) is a measure of the magnitude of an effect, and thus may indicate the degree of practical importance of our results. Note that these two concepts are, in principle at least, orthogonal to each other. As Rosnow and Rosenthal (1989) have pointed out: "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" (p. 1279). Thus, statistical significance and effect size are both important because they provide complementary information about reliability and magnitude, respectively.

However, in an applied science like human factors engineering, effect size plays a particularly important role. As Chow (1996) has 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" (p. 8). Despite this truism, even just a cursory examination of the human factors literature reveals that statistical significance is reported far more frequently than is effect size. Once again, we believe that this is indicative of an over-reliance on NHST and ANOVA. Neither of these statistical techniques provides direct measures of effect size. Instead, their focus is on reliability.

Because of the foundational importance of practical significance to human factors engineering, it is important that we calculate effect sizes in addition to assessing statistical reliability. Several ways of calculating effect size have been proposed in the literature. For

example, Cohen (1988) has proposed the standardised mean difference statistic, *d*, as a generalisable 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 reliability and magnitude of an effect is best conveyed by an example (adapted from Rosnow & Rosenthal, 1989). Consider two hypothetical experiments, both conducted to evaluate the impact of two types of training programs, *T*1 and *T*2, on human performance. In one experiment (with $n = 80$), T1 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 NHST, 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 true value of *T*1 on human performance.

Calculating effect size adds new information that can help put the results in a more realistic light. In our 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 reliability (indicated by statistical significance) and magnitude (indicated by effect size), and thus, why it is important to calculate effect size.

Rouanet (1996) has proposed a set of Bayesian statistics that extend the usefulness of effect sizes. Rouanet's technique allows us to make inferences about how large or how small an effect is in a population. In the following section, we provide an example of how this analysis technique can be used to assess the practical significance of human factors research.

In summary, the emphasis on NHST and ANOVA has led to an emphasis on statistical reliability to the detriment of an emphasis on effect magnitude. Cohen's (1988) *d* and Rouanet's (1996) Bayesian extensions of this statistic provide rigorous and systematic ways of calculating

effect size, thereby allowing human factors researchers to assess the practical significance of their findings.

5. The null hypothesis is never true.

In addition to the points discussed so far, there is another reason for not putting too much emphasis in 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 never really true in behavioural research. This point has been made by many noted researchers (Abelson, 1995; Cohen, 1990, 1994; Loftus, 1991, in press; Meehl, 1967, 1978, 1990, 1997; Steiger & Fouladi, 1997), but as with the other points we have discussed, its implications have not been taken as seriously as they should be.

Consider a typical human factors experiment comparing the effect of two treatments (e.g., two interfaces, two training programs, or two selection criteria) on human performance. One group of subjects 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 to point out that the answers to questions such as this one are 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. (Meehl, 1967, p. 162)

As we mentioned earlier, Meehl is not alone in his opinion. Many other noted researchers have voiced the same criticism.

One way to illustrate the unrealistic 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 $p < 0.000001$.

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, if we have a sample size of 14,000, then 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.

If the null hypothesis is always false, then the act of conducting a NHST means something very different than what we usually think 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 we have the fortitude and resources to include enough subjects in our experiments, then virtually any null hypothesis can be rejected. Thus, the value of just conducting a NHST is minimal. As Cohen (1994) has pointed out, "if all we ... learn from a research is that A is larger than B ($p < .01$), we have not learned very much. And this is typically all we learn" (p. 1001).

If we accept the fact that the null hypothesis is never true in behavioural research, what are the implications for the statistical analysis of data? The short answer to this question 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 (Abelson, 1995; Cohen, 1990, 1994; Loftus, 1993b, 1995, in press; Loftus & Masson, 1994; Meehl, 1997; Steiger & 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. The resulting measure provides an explicit indication of the sensitivity of an experiment to detect an effect of interest. The calculation of power is especially valuable in cases 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 researchers from "falsifying" theories simply by not including enough subjects in their

experiment, it would be useful to present calculations of power. Doing so would provide additional information than 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 a pattern of population parameters. 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. Second, the width of a confidence interval provides an indication of the precision of measurement (or power). 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. Third, the relative position of two or more confidence intervals can provide 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, otherwise they are not. Finally, confidence intervals also provide the same information that is usually obtained from a NHST. If the confidence interval includes the value zero, then the NHST is not significant, otherwise it is. Therefore, the plotting of confidence intervals provides researchers with more insights into their data than could be obtained by NHST alone.

The informativeness of confidence intervals can be illustrated with a simple example borrowed from Steiger and Fouladi (1997). Figure 3 shows confidence intervals for the differences between means from three hypothetical experiments. 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 reliably different from zero, leading to a statistically significant decision to reject the null hypothesis. In Experiment 2, the confidence interval does include zero. Thus, in this case, a NHST would indicate that the difference in means is not reliably different from zero. Thus, the confidence intervals Figure 3 provides the information that can be obtained directly from a NHST with the difference that that information is presented graphically.

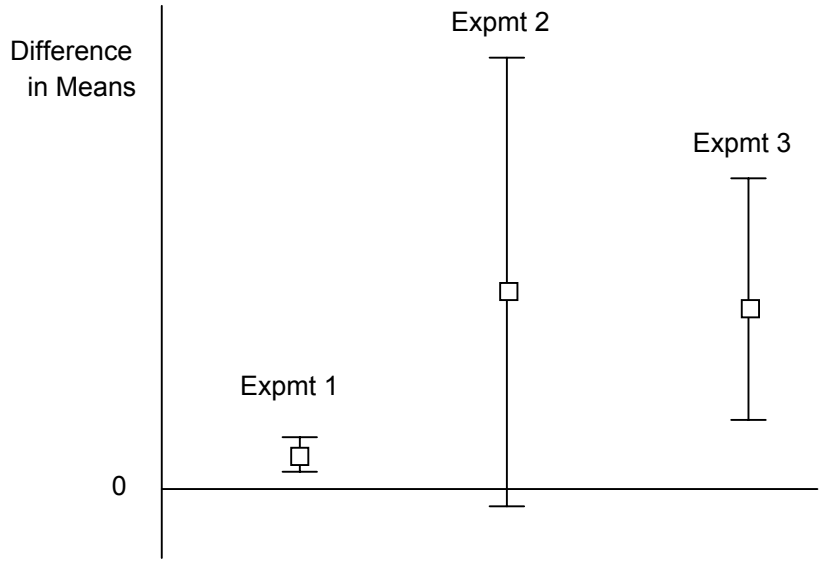

Figure 3: Hypothetical example showing how confidence intervals reflect different degrees of measurement precision (adapted from Steiger & Fouladi, 1997).

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 not agree with those from the other two experiments. The confidence intervals provide a graphical basis for reaching a very 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 which 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 measurement precision was just inadequate to detect that effect. Thus, the results from Experiment 2 are imprecise, and thus we do not know 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, which indicates 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 the hypothetical example in Figure 3 is that confidence intervals provide much more information than do NHSTs. Furthermore, that information is provided in a graphical format, thereby making it easier for researchers 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 lead to a very different interpretation than would 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 our data, and thereby allow us to go beyond merely rejecting the null hypothesis. Note that confidence intervals can be calculated for effect sizes as well, thereby combining several of the advantages of each of these techniques into one statistical procedure (Cohen, 1990, 1994; Rosnow & Rosenthal, 1989). In this way, we would obtain information about the reliability of our knowledge of effect size, information that is surely to be of practical value in human factors research.

6. One experiment is always inconclusive.

This final point is deeper than the others in that it cuts across the comparative advantages and disadvantages of any particular set of statistical analysis techniques. No matter how carefully it is designed, not matter how sophisticated the equipment, and no matter what statistical analysis techniques are used, any one experiment 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 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, ten 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 own 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, p. 253). In short, there is no such thing as a "critical experiment" because empirical knowledge is inductive, and thus, fragile when viewed in isolation. Like the other points that we 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 (Cohen, 1990; Dar, 1987; Rosnow & Rosenthal, 1989; Rossi, 1997; Schmidt & Hunter, 1997), the way in which NHST and NOVA 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, "significance tests cannot separate real findings from chance findings in research studies" (Schmidt & Hunter, 1997, p. 39).

Researchers frequently ignore the fact that there is no objective, mechanical procedure for making a dichotomous decision to evaluate the validity of research findings. 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 investigation (1948-1969), about 40 papers 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 pretty much abandoned.

Rossi (1997) conducted a retrospective analysis of the collective findings in this body of literature. Data from a total of 47 experiments with an aggregate of 4,926 subjects were included in the analysis. Of these studies, only 43% reported statistically significant results at $p < 0.05$. It

is this low percentage of significant results which led researchers to doubt the existence of the spontaneous recovery effect. However, when the experiments were analysed as a whole, there was very reliable evidence in support of the spontaneous recovery effect ($p < 0.001$). Rossi also conducted an effect size analysis and a power size 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 results help 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. But as Rosnow and Rosenthal (1989) have observed, "dichotomous significance testing has no ontological basis .... surely, God loves the .06 nearly as much as the .05" (p. 1277). Because of the combination of small effect and low power, 57% of the experiments did not generate results that passed the naive 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 we conclude 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. Second, 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) 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" (p. 183). Third, 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 to building

scientific knowledge. In the words of Abelson (1995), "Research conclusions arise not from single studies alone, but from cumulative replication" (p. 77). Even if no single result reaches statistical significance at the $p < 0.05$ value, the entire pattern of results can be highly reliable 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, p. 1311).

How many other cases like the one reviewed by Rossi (1997) are there in the literature? It is very difficult to answer this question. Nevertheless, there is one thing that we can be sure of. Making decisions on a dichotomous basis using NHST will only make it more likely for such problems to plague the 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., Meehl, 1997; Rosnow & Rosenthal, 1989; Schmidt & Hunter, 1997). This lesson is perhaps the most important one of all among the ones that we have reviewed.

APPLICATION TO PREVIOUS JAERI EXPERIMENTS

Introduction

In this section, we will apply some of the statistical analysis techniques identified in the previous section to data from previous experiments conducted for JAERI. We will begin by introducing the techniques that we have chosen to apply, and will then outline the theory behind each technique and a methodology for applying them to the JAERI dataset. Results will be presented and discussed, and conclusions both about the research programme as a whole and about the usefulness of these novel methods will be made.

Techniques Applied

Several lesser known statistical methods were introduced in our literature review. These methods were: (1) individual subjects analyses using non-parametric statistics (Hammond et al., 1987; Vicente, 1992), (2) Campbell & Fiske's (1959) Multitrait-Multimethod matrix and extensions of this technique (Hammond et al., 1986), (3) confidence intervals (Loftus, 1993b; Loftus & Masson, 1994), (4) power analysis (Cohen, 1988; Kirk, 1995; Pearson & Hartley, 1951), (5) effect size evaluation (Cohen, 1994), and (6) the use of Bayesian methods to assert largeness or smallness of effects (Rouanet, 1996). In this section, we will apply three of these

techniques (confidence interval estimation, power analysis, and Bayesian methods to assert largeness or smallness of effects) to the JAERI dataset. Since one dataset cannot possibly afford the full range of statistical techniques presented, these techniques were chosen for the sole reason that they are well suited to our data. This is not to say that the other techniques are not useful to this programme of research, but rather that their usefulness is promissory.

Confidence Intervals and Graphical Analyses of Variance

As already discussed, Loftus (1993b) and Loftus and Masson (1994) have written about the drawbacks of traditional NHST and promote the practice of confidence interval (CI) estimation to overcome many of these drawbacks and to aid in data description. To review, while NHST techniques output a probability that some set of population means is not equal to some other set, they provide no direct information about the power of this conclusion, the actual relationship between the means, or the size of effects. CI estimation, on the other hand, provides a richer set of information. CIs are able to portray a standard NHST graphically, while also providing information to allow an experimenter to intuitively assess the relationship between the means, the power of the conclusion, and the size of effects.

While Loftus and Masson (1994) describe CI estimation for both between- and within-subject experiments, since the JAERI experiments primarily involved between-subjects designs, only these techniques will be covered here. The reader is referred to Loftus and Masson (1994) for a discussion of extensions of these techniques to within-subjects designs.

Most standard introductory statistical textbooks cover the topic of CI estimation (though, as Loftus and Masson (1994) observe, the importance of this technique is not stressed). A standard CI about the mean for the case in which the means and standard deviations are not known, but in which the sampling distribution of the means can reasonably be assumed to be normal, is based on the *t* distribution[1]. Given a type I error probability of α, a two-sided 1-α CI is given by:

$$\left\{ \overline{X} - \frac{t_{\alpha/2,v}\hat{\sigma}}{\sqrt{n}} < \mu < \overline{X} + \frac{t_{\alpha/2,v}\hat{\sigma}}{\sqrt{n}} \right\} \qquad (1)$$

where

$\overline{X}$ is an estimate of the population mean,

[1] For an extended discussion of simple CI estimation, see Hines & Montgomery (1990).

$\hat{\sigma}$ is an estimate of the population standard deviation,

n is the sample size, and

ν is the error degrees of freedom, which in this simple case is equal to n - 1.

Loftus and Masson (1994) have extended this procedure to use information contained in any standard ANOVA table. Since the expected value of the mean squared error, $E(MS_E)$, is the population standard deviation, σ_ε^2, Equation 1 can be rewritten as:

$$\left\{ \overline{X} - t_{\alpha/2,\varphi} \sqrt{MS_\varepsilon / n} < \mu < \overline{X} + t_{\alpha/2,\varphi} \sqrt{MS_\varepsilon / n} \right\} \quad (2)$$

where

$\overline{X}$ and n are the same as in equation 1, and

φ is the error degrees of freedom, from the ANOVA table.

Or, in a more tractable format:

$$CI = M_j \pm \left(\sqrt{MS_\varepsilon / n} \right) t_{\alpha/2,\varphi} \quad (3)$$

Using this formula, CIs for all of the j groups included in an experimental design can be calculated and compared against one another. Standard hypothesis testing can be done by comparing the CIs of the different experimental groups. If the CIs from two groups do not overlap, the null hypothesis that the two means are equal is rejected; conversely, if the CIs do overlap, then we fail to reject the null hypothesis. While this procedure does not output a quantitative p value, Loftus and Masson (1994) and the authors do not view this as a drawback. Rather, one of the real advantages of using CIs to test hypotheses (in other words, a graphical ANOVA) is that it takes the focus away from arbitrary p-values and places the focus on the pattern of means and the variability in the data. The use of CIs helps to shift the product of data analysis from parametrization (lists of p-values) to description (an intuitive understanding of the variability in the data set and of the relationships between the sets of means).

Power Analysis

Power analysis is the formal determination of the sensitivity of a statistical test (and by implication, of an experiment) to detect differences in a set of dependent variables based on the manipulation of one or more independent variables. Although theoretically quite a complex

procedure,[2] the actual application and interpretation of power analysis for most ANOVA designs is quite simple.

In most applications of ANOVA, researchers tend to be most concerned with ensuring that they do not commit the error of rejecting the null hypothesis if it is, in fact, true. This is a Type I error, and the probability of it occurring is explicitly set at α (the significance level of the test). In many situations α is set at .05 or less as Type I errors are viewed as especially grave. Type II errors are committed when there is a failure to reject the null hypothesis when it is, in fact, false. The probability of a Type II error, β, cannot be set to a certain value by the experimenter in the same way as α can be, but rather is a function of both the sample size and the population effect size. Strictly speaking, β can only be determined after an experiment has been performed and an estimate of the population effect size is known.

The power of a test is equal to 1 - β, and is the probability that a false null hypothesis is correctly rejected. While no strict guidelines exist for how large the power of the test should be (perhaps because few tests achieve even moderate power), it is generally acknowledged that a power of greater than .8 is desirable for an experimental test (Cohen, 1988). The power of an ANOVA can be determined by first calculating the value of ϕ:

$$\phi = \sqrt{\frac{\sum_{j=1}^{p} \alpha_j^2 / p}{\sigma_\varepsilon^2 / n}} \tag{4}$$

where:

α_j is the size of treatment effect j,

p is the number of treatments,

σ_ε^2 is the standard deviation of error effects, and

n is the sample size.

To simplify calculations,

$$\frac{\sum_{j=1}^{p} \alpha_j^2}{p} = \frac{p-1}{np}(MS_{Error} - MS_{Effect}) \text{ and } \sigma_\varepsilon^2 = MS_{Error} \tag{5}$$

[2] For an excellent treatment of power analysis, see Cohen (1988).

Once ϕ has been calculated it can be compared to charts of the power function for ANOVA (see Kirk, 1995; Pearson & Hartley, 1951), which are based on four parameters: ν_1, ν_2, the chosen significance level α, and ϕ, where $\nu_1 = p - 1$ and $\nu_2 = p(n-1)$. To use these charts, the chart for the appropriate ν_1 is located, on which power can be determined for a given ϕ, ν_2, and α (generally either .05 or .01).

Two issues should be noted. First of all, power cannot be calculated for cases in which $MS_{Error} > MS_{Effect}$, that is where $F < 1$. These cases will result in a negative value for $\sum_{j=1}^{p} \alpha_j^2 \Big/ p$, and will make ϕ undefined. Power is very low in these cases (less than 0.30). Second, and more importantly, the method outlined above is used to calculate power post-hoc and so outputs the power of the experiment *given the data observed*. If power is defined as the ability of an experiment to detect effects, then highly significant effects imply high power. If significant effects were not observed, power will generally be low. This can be better understood by looking at the formula for ϕ. As the difference between MS_{effect} and MS_{error} grows in favour of MS_{effect}, both F and $\sum_{j=1}^{p} \alpha_j^2 \Big/ p$ will become large. All things being equal, as F increases, p decreases, and as $\sum_{j=1}^{p} \alpha_j^2 \Big/ p$ increases, power increases (see equation 4).

Bayesian Methods to Assert Largeness or Smallness

Of the three types of analyses presented in this section, the Bayesian methods that follow will be the most unfamiliar to many readers. This is because Bayesian techniques differ greatly from typical ANOVA analyses. While ANOVA techniques ask the oblique question, "could the observed set of data have occurred if the null hypothesis is true?", Bayesian techniques allow one to ask the more direct question, "what is the probability that this hypothesis is true?" In spite of this direct approach to statistical inference, the reason that some practitioners have been hesitant to adopt Bayesian techniques is that they generally require one to first make an estimate of the probability that the hypothesis is true (the *prior* probability) in order to come up with a refined estimate of how much belief to put in the hypothesis in light of the data (the *posterior* probability). Stating a prior probability before carrying out an investigation is seen by some as unobjective and unscientific.

In light of this criticism, Rouanet (1996) has presented a technique that is important in two aspects. First of all, it makes the problem of specifying a prior probability moot, as it uses a prior that expresses ignorance about the parameters of interest. In this way, the posterior expresses only evidence from the data and is not coloured by the experimenters own prior beliefs. Second, the purpose of this method is not to test standard null hypotheses about the presence of an effect, but rather to test hypotheses about the size of the observed effect. It allows the calculation of both observed effect sizes and confidence intervals about these observations. This is useful for a number of reasons. First of all, observed effect sizes are independent of statistical significance. In experiments that may not achieve statistical significance, Bayesian methods still allow us to make strong assertions about how large (or small) an effect could reasonably be. Second, the investigation of effect sizes can be very helpful in describing the observed data. Effect sizes can help in understanding the central tendency in the data while a confidence interval about the effect size can help in understanding the degree of belief that can be placed in this point estimate.

A method of assessing largeness or smallness of effect size is described below. This discussion closely follows that of Rouanet (1996), to which the reader is referred for a more complete development.

Effect sizes. Consider an investigation where n subjects undergo two treatments, each resulting in some score. If d_i is the difference between the two scores for subject i, then the mean difference is $\sum_{i=1}^{n} d_i / n = \bar{d}$, and will be referred to as d from now on. d is an estimator of δ, the actual population effect. Similarly, $s^2 = \sum_{i=1}^{n} \frac{(d_i - d)^2}{(n-1)}$ is an unbiased estimator of σ^2 (with $q = n - 1$ *df*), the variance of individual effects in the population. Having established these conventions, we will from now on refer to the *standardised effect*, $|d|/s$ (see Cohen, 1988). By convention, $|d|/s \cong .5$ is considered a medium-sized effect, $|d|/s < .4$ is considered a small effect, and $|d|/s > .6$ is considered a large effect. While these values are only rules of thumb, they do specify a benchmark for a starting point.

Assumptions. Now assume that the d_i's are independent and normally distributed, $N(\delta, \sigma^2)$. Then the sampling distribution of the mean d is $N(\delta, \sigma^2/n)$. To test the null

hypothesis $H_0 : \delta = 0$, the usual test statistic is $t = \sqrt{n}(d / s)$, which under H_0 is distributed as an elementary t with $q = n - 1$ *df*.

Prior distribution. At this point, Bayesian elements are introduced into the analysis. The typical Bayesian approach is to postulate some prior distribution that expresses one's certainty about the parameters of interest independently of the data at hand, and to combine this prior with the sampling distribution to yield a posterior distribution that expresses uncertainty about the parameters conditional on the new data. Instead of expressing any certainty about the parameters in the prior, Rouanet (1996) advocates the use of a *noninformative* prior that expresses a state of ignorance about the parameters. In this way, the posterior will express only the evidence brought by the data.

Posterior distribution. Assuming this noninformative prior on (δ, σ) and given $d = d_{obs}$ (the observed value of d) and $s = s_{obs}$ (the observed value of s), the posterior distribution of δ is such that $\sqrt{n}(\delta - d_{obs}) / s_{obs}$ is distributed as an elementary t variable. The distribution of δ is then a scaled t with mean d and scale parameter s_{obs}^2 / n, or in more formal notation,

$$\delta \sim t_q\left(d_{obs}, \frac{s_{obs}^2}{n}\right). \qquad (6)$$

Of import for drawing inferences about effect sizes is the connection between the sampling and posterior distributions, which can be seen by scaling both d_{obs} and δ by s_{obs}. If we let $(d / s)_{obs}$ denote the observed standardised effect, d_{obs} / s_{obs} then the posterior distribution $\delta / s_{obs} \sim t_q[(d / s)_{obs}, 1/n]$ can be found by shifting the sampling distribution under H_0 by $(d / s)_{obs}$.

Methodology. To use the above development to draw inferences about effect sizes, the observed effect is taken as an estimate of the population effect, and the approach outlined constructs a distribution around this estimate. In effect, this type of inference attempts to understand what a population that has produced the observed effect might look like. The conclusions that can be drawn from a posterior distribution supplement the descriptive conclusions that can be made from effect sizes. If most of the posterior distribution lies in the area of large effects, then there is a great probability that the effect is indeed large. On the other

hand, if most of the posterior distribution lies in the area of small effects, there is a great probability that the effect is small. So, using this method, assertions can be made about the largeness or smallness of an effect.

Examples. Rouanet (1996) presented two examples which are helpful to reproduce here. These examples have been fleshed out somewhat to better describe the calculations involved. Both of these examples involve four parameters: (1) a *credibility level* γ (> 0.5) which expresses the level of certainty than any assertion will have to have, (2) a *limit for largeness,* l_{lar} or *limit for smallness,* l_{sma}, which express how large or how small an effect must be to count as either large or small, respectively, (3) $(d/s)_{obs}$, the observed effect size, and (4) the sample size, n.

In the first example, consider an experiment where $(d/s)_{obs} = 0.9$ with $n = 25$. The posterior distribution is then $\delta / s_{obs} \sim t_{24}(0.9, 1/25)$. To see whether or not largeness can be asserted in this case, we set the credibility level, γ, at 0.9, and the limit for largeness, l_{lar}, at 0.6, and then attempt to determine if $P(\delta / s_{obs} > l_{lar}) > \gamma$. To do this, the posterior distribution must be shifted to the elementary t distribution[3]:

$$P(\delta / s_{obs} > l_{lar}) = P\left(t_q > \frac{x - \mu_t}{\sigma_t} \right) \qquad (7)$$

where:

x = l_{lar},

$\mu_t = (d/s)_{obs}$, and

$\sigma_t = 1/\sqrt{n}$.

Then,

$$P(\delta / s_{obs} > 0.6) = P\left(t_{24} > \frac{0.6 - 0.9}{1/5} \right)$$

$$= P(t_{24} > -1.5)$$

$$= 0.927.$$

Since $P(\delta / s_{obs} > 0.6) = 0.927 > 0.9$, we can assert largeness in this case.

[3] For more information on the mechanics of shifting a *t*-distribution, see Phillips (1973).

As a second example, consider an experiment where $(d/s)_{obs} = 0.1$ with $n = 25$. The posterior distribution is then $\delta / s_{obs} \sim t_{24}(0.1, 1/25)$. To see whether or not smallness can be asserted in this case, we set the credibility level, γ, at 0.9, and the limit for smallness, l_{sma}, at 0.4, and then attempt to determine if $P(\delta / s_{obs} < l_{lar}) > \gamma$. To do this, we again shift the posterior distribution to the elementary *t*:

$$\begin{aligned} P(\delta / s_{obs} < 0.4) &= P\left(t_q < \frac{x - \mu_t}{\sigma_t} \right) \\ &= P\left(t_{24} < \frac{0.4 - 0.1}{1/5} \right) \\ &= P(t_{24} < 1.5) \\ &= 0.927. \end{aligned}$$

Since $P(\delta / s_{obs} < 0.4) = 0.927 > 0.9$, we can assert smallness in this case.

Extensions to ANOVA. These procedures can easily be extended to complement the inferences made using an ANOVA. While Rouanet (1996) has extended these procedures to include all ANOVA inferences, inferences on effects with 1 *df* are treated differently than those with $df > 1$. Only inferences for 1 *df* will be treated in this paper as we were not able to obtain the software necessary to perform the multi-*df* analyses.

The posterior distribution used to make inferences on 1 *df* ANOVA effects is similar to that presented earlier:

$$\frac{\delta}{s_{obs}} \sim t_q \left[\left(\frac{d}{s} \right)_{obs}, \left(\frac{1}{n} \right) \right] \qquad (8)$$

As a result, Bayesian inference for any 1 *df* source of variation involves only the *t* distribution. Since 1 *df* sources of variation are so common in experimental practice, it is fortunate that such a straightforward posterior distribution exists. Calculations for inferences on 1 *df* sources of variation are most easily made by reference to data in the ANOVA table. How this is done is best demonstrated by means of an example.

Example: Effect sizes for interface effect in JAERI II[4] for diagnosis score. In the JAERI II investigation, fault diagnosis was coded on a scale of 0 to 3 (Pawlak & Vicente, 1996). Subjects who did not notice the fault would be assigned a score of 0 for that fault, while those who isolated the fault and its root cause would be assigned a 3. A partial ANOVA table to test hypotheses on this measure is reproduced in Table 3.

Table 3: Partial ANOVA table for JAERI II diagnosis score.

Source	df	SS	MS	F	p
INT	1	38.128	38.128	15.01	0.0009
TRAIN	1	0.832	0.832	0.33	0.5734
INT*TRAIN	1	4.730	4.730	1.86	0.1875
SUBJECT(INT*TRAIN)	20	50.803	2.540		

To find the effect size for INT, four parameters are needed: the effect mean square (MS_{eff}), the error mean square (MS_{err}), the number of observations for every level of the effect (n_{eff}), and the number of subjects per group (n_{err}). In this case, $MS_{eff} = 38.128$, $MS_{err} = 2.540$, $n_{eff} = 129$ (the number of faults for which we have data per group), and $n_{err} = 12$. Using these data, $\hat{n}$, s^2_{eff}, and s^2_{err} can be found, as can $|d/s|_{obs}$:

$$\hat{n} = n_{eff} / n_{err} \tag{9}$$

$$s^2_{eff} = MS_{eff} / n_{eff} \tag{10}$$

$$s^2_{err} = MS_{err} / n_{err} \tag{11}$$

$$|d/s|_{obs} = \sqrt{s^2_{eff} / s^2_{err}} \tag{12}$$

In this case, $\hat{n} = 12$, $s^2_{eff} = 0.29556$, $s^2_{err} = 0.21167$, and $|d/s|_{obs} = 1.182$. With this information, the posterior distribution can be constructed and inferences can be made. Using the form of the posterior distribution in equation 8,

$$\frac{\delta}{s_{obs}} \sim t_q\left[\left(\frac{d}{s}\right)_{obs}, \left(\frac{1}{\hat{n}}\right)\right]$$

[4] Familiarity with the JAERI II experiment is not necessary to understand this example. Readers who are not familiar with the programme of experiments carried out for JAERI and are interested in understand the context of this example are referred to the section titled **Background** in which the JAERI II experiment is briefly introduced.

where q is equal to the error *df* from the ANOVA table. So the posterior distribution for this example is:

$$\frac{\delta}{s_{obs}} \sim t_{20}\left[1.182, 1/12\right].$$

Now that the posterior distribution is known, we can proceed as in the previous examples to see if largeness can be asserted in this case. Using $l_{lar} = 0.6$ and $\gamma = 0.9$,

$$\begin{aligned} P(\delta / s_{obs} > 0.6) &= P\left(t_q > \frac{x - \mu_t}{\sigma_t}\right) \\ &= P\left(t_{20} > \frac{0.6 - 1.182}{1/\sqrt{12}}\right) \\ &= P(t_{20} > -1.794) \\ &= 0.971. \end{aligned}$$

Since $P(\delta / s_{obs} > 0.6) = 0.971 > 0.9$, we can assert largeness in this case. Therefore, we can say with a relatively high degree of certainty that the effect of interface on diagnosis score in the JAERI II experiment was large. Performing similar calculations for the TRAIN effect, it can be found that the posterior distribution for this effect is:

$$\frac{\delta}{s_{obs}} \sim t_{20}\left[0.1746, 1/12\right].$$

Taking $l_{sma} = 0.4$ and $\gamma = 0.9$, $P(\delta / s_{obs} < 0.4) = 0.778 < 0.9$. As this is the case, we cannot assert smallness here. While the *p*-value from the ANOVA table allows us to conclude little more than that the effect is not significant (remember, large *p*-values do not allow us to conclude that no effect exists), this Bayesian procedure allows us to conclude that the true effect is most probably not small.

Bayesian confidence intervals. Rouanet's (1996) procedures can be easily extended to a graphical presentation in the form of confidence intervals around $|d / s|_{obs}$. These confidence intervals present information about both the magnitude of $|d / s|_{obs}$ and the degree of certainty that can be put in this measure. It should be stressed, however, that the power of an experiment cannot be ascertained from CIs around $|d / s|_{obs}$ in the same way that this can be done for CIs around means. The width of a CIs around $|d / s|_{obs}$ is strongly determined by the sample size, larger samples resulting in smaller CIs. The width of a CI around the mean, on the other hand, is

strongly determined by both the sample size and the variability in the data – large sample sizes and low variability in tandem result in smaller CIs. Thus, confidence intervals around $|d/s|_{obs}$ serve only to show boundaries for asserting either largeness or smallness at a given level of γ.

The logic behind constructing a confidence interval around $|d/s|_{obs}$ is straightforward. Instead of using the posterior distribution to see if a given effect size can reasonably be called either small or large at a level of confidence γ, the posterior distribution is used to infer what the upper and lower γ limits are. So, to find the upper limit of the CI, we need to find for what effect size

$$P(\delta / s_{obs} > \beta_{upper}) = \gamma$$

To do this, set:

$$\gamma = t_q\left(\frac{\beta_{upper} - (d/s)_{obs}}{1/\sqrt{n}}\right)$$

and solve for β:

$$t_q^{-1}(\gamma) = \frac{\beta_{upper} - (d/s)_{obs}}{1/\sqrt{n}}$$

$$\beta_{upper} = \frac{t_q^{-1}(\gamma)}{\sqrt{n}} + (d/s)_{obs}. \qquad (13)$$

While finding the lower limit (β_{lower})can be done in the same manner, since the *t* distribution is symmetrical, the upper limit can be mirrored to achieve a two-sided $(1-2\gamma)$ confidence interval on the effect size (i.e., if in the above calculations γ were set to .9, this procedure would result in an 80% CI around $|d/s|_{obs}$). Using the data for the INT effect example presented earlier, with γ = 0.9,

$$\beta_{upper} = \frac{t_{20}^{-1}(0.9)}{\sqrt{12}} + 1.118$$

$$\beta_{upper} = 1.5$$

$$\beta_{lower} = 2(d/s)_{obs} - \beta_{upper}$$

$$\beta_{lower} = .736.$$

Similarly, for the train effect, with γ = 0.9,

$$\beta_{upper} = \frac{t_{20}^{-1}(0.9)}{\sqrt{12}} + 0.1652$$

$$\beta_{upper} = 0.5477$$

$$\beta_{lower} = (2)(0.1652) - 0.5477$$

$$\beta_{lower} = -0.3304$$

Since negative effect sizes are not possible, β_{lower} is set at 0. With this information, the confidence intervals around the effect sizes for both the INT and TRAIN can be plotted (see Figure 4). Notice that assertions about smallness and largeness can be read directly from this figure. Since the confidence interval for the INT effect does not drop below 0.6, we can assert with at least a 90% level of confidence that this effect is large. Similarly, since the confidence interval for the TRAIN effect stretches above 0.4, we cannot assert smallness with a 90% level of confidence.

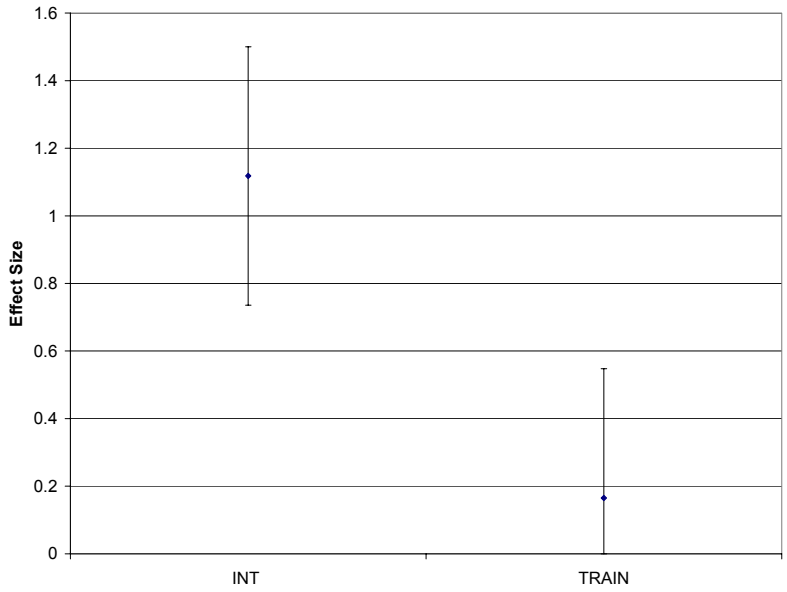

Figure 4: 80% CIs about effect sizes for JAERI II trial completion time, by effect.

Results

Background

To set the stage for the analyses that follow, we will first summarise the data from the four studies that comprise the database for these analyses. All of these studies have been conducted in the context of the DURESS II testbed, a simulation of a highly simplified yet representative

thermal hydraulic plant. DURESS II can be controlled using either a conventional interface displaying only physical information (P interface) or an interface developed according to the principles of EID that displays both physical and functional information (P+F interface).

Using DURESS II, each experiment was designed to investigate the effects on operator adaptation of modifying one or more behaviour shaping constraints. Generally speaking, behaviour shaping constraints are any type of constraint that may shape how operators adapt to a work domain. The specific behaviour shaping constraints relevant to this program of research are interface content, interface form, type of training, and pre-existing competencies (Howie, Janzen, & Vicente, 1996):

- **Interface Content.** Interface content is a strong constraint on operator performance (Christoffersen et al., 1994). While providing proper and relevant information is a necessary (but not sufficient) provision for functional adaptation, either neglecting to include critical information or providing irrelevant information can foster dysfunctional adaptation.

- **Interface Form.** Independent of the information content of an interface is the form of information presentation. Operators may become increasingly attuned to the visual form of an interface, or in other words, the visual form of an interface will manipulate an operator's attention. An interface that directs an operator's attention to critical information should foster functional adaptation. Conversely, an interface that directs an operator's attention to either non-critical or irrelevant information may promote dysfunctional adaptation.

- **Type of Training.** The type and amount of training that operators receive influences adaptation. Operators can receive training either prior to or concurrent with operating a system. Training provides operators with: (1) a set of competencies tailored to a specific work situation, (2) guidance in what types of information to treat as important, and (3) experience for dealing with novel situations. Many types of training exist, and some research (e.g. Crossman & Cooke, 1962/1974) indicates that at least some types of theoretical training do not foster functional adaptation. The effect of training based on both fundamental physical principles and interface design, however, could foster functional adaptation.

- **Pre-existing competencies.** The subjects in each experiment have pre-existing competencies that influence adaptation. These take on such forms as cognitive style, declarative knowledge, perceptual and motor skills, and population stereotypes. Although this set of behaviour shaping constraints cannot be controlled in the same fashion as those listed above, an understanding of their effects is important for both experimental design and analysis of results.

Table 4 summarises the manipulations of behaviour shaping constraints across a series of four experiments. The accompanying text describes each of these experiments.

Table 4: Integrated summary of the three-year research program, showing the behaviour shaping constraints investigated in each experiment. Adapted from Howie et al. (1996 p. 9).

	Behaviour Shaping Constraints				
Experiment	**Interface Content**	**Interface Form**	**Type of Training**	**Pre-existing competencies**	**Number of Subjects**
JAERI I	P vs. P+F	P vs. P+F	None	Demographic Data + Cognitive Style	6
JAERI II	P vs. P+F	P vs. P+F	None vs. AH		24
JAERI IIIa	P+F	P vs. P+F vs. Divided P+F	None		12*
JAERI IIIb	P+F	P+F	None vs. Dplayer vs. Dplayer + SE		18*

* JAERI IIIa and IIIb made use of a shared control group of 6 subjects. In total, 24 subjects participated in JAERI IIIa and IIIb.

1. **JAERI I** (Christoffersen et al., 1994): This experiment was designed to assess the impact of interface content and form on long-term adaptation. It involved a longitudinal investigation in which six subjects operated either the P or the P+F interface of the DURESS II microworld quasi-daily over a period of six months (217 trials).

2. **JAERI II** (Hunter, Janzen, & Vicente, 1995): This experiment investigated the interaction between interface design and model-based training on adaptation. Twenty-four subjects participated in a 2 x 2 experiment, with two levels of interface design (P vs. P+F) and two levels of training (none vs. abstraction hierarchy [AH] training), over a period of one month (67 trials).

3. **JAERI IIIa** (Howie et al., 1996): This experiment investigated the impact of interface form on adaptation. The P+F interface of DURESS II was compared to a divided P+F interface with one level of information for each level of the abstraction hierarchy. 12 subjects, divided

into two groups (integrated vs. divided), participated in this experiment for one month (67 trials) each.

4. **JAERI IIIb** (Howie et al., 1996): This experiment investigated the effect of a second type of training, self instruction via performance reviews and/or self-explanation, on operator performance. 18 subjects were divided into three groups, each of which performed identical tasks on the P+F interface while engaging in different levels of performance reviews and/or self-explanation. The first group did not review their performance or engage in any self-explanation of control actions. The second group periodically reviewed their performance using the Dplayer program, a program that plays back trials in real-time from data contained in the simulator log files. The third group also periodically reviewed their performance using Dplayer, but its members were also instructed and encouraged to engage in self-explanation of control actions while reviewing their trials.

In each of these experiments, subjects would perform various tasks on DURESS II under both normal and fault conditions. In the analyses that follow, six measures used to gauge subject performance will be referred to. These are:

1. **Trial Completion Time (TCT)**, the time taken for subjects to complete normal trials.
2. **Trial Completion Time Variance (CTV)**, the variance in completion times over a block of trials.
3. **Fault Detection Time (DET)**, the time elapsed between the occurrence of a fault and the subjects' verbal detection of that fault.
4. **Diagnosis Accuracy (DA)**, a score assigned to subjects' diagnoses, ranging from a score of 0 for an irrelevant utterance to 3 for a statement of the location and root cause of the fault (Pawlak & Vicente, 1996).
5. **Diagnosis Time (DGT)**, the time elapsed between the occurrence of a fault and the subject's verbalization of a correct root cause diagnosis (i.e., DA of 3).
6. **Compensation Time (CT)**, the time elapsed between the onset of a fault and a subject's proper termination of a trial. Trials for which subjects were not able to regain control over the plant after a fault are treated as missing data.

While previous analyses of the data from these experiments have been thorough and have uncovered many interesting results, the three techniques described above have not yet been applied to these data. This will be done in the sections that follow.

Confidence Intervals and Graphical Analyses of Variance

In the introduction to CIs and graphical analyses of variance, it was stressed that this technique is a useful way of using the information in an ANOVA table for data description. Since we will be doing explicit analyses of power and effect size in later sections, graphical ANOVAs will not help us to learn a great deal that is new in the context of these already completed investigations. Rather, they will help to make future analyses more informative and efficient. As this is the case, in this section we will not construct graphical ANOVAs to retell the stories of old analyses. This section will simply serve as a further introduction to the use of CIs through two concrete and informative examples taken from the JAERI dataset.

These two cases both involve the JAERI II investigation (Hunter et al., 1995), and involve the construction of CIs around the group means of trial completion time for the two main effects of the investigation, interface and training. In the original contract report, Hunter et al. (1995) reported that for the dependent variable TCT, there was a significant interface effect but no significant training effect. While means for the interface condition were reported (revealing that subjects using the P interface had faster completion times than those using the P+F interface), means for the training condition were not reported. This leaves at least four relevant questions unanswered: (1) In general terms, how powerful is the conclusion on the interface (INT) effect? (2) How close to significant (i.e., $p < .05$) was the training (TRAIN) effect? (3) What was the pattern of observed means for the two effects? (4) What are the effect sizes for both INT and TRAIN? Graphical ANOVAs will be constructed for both of these effects to answer these questions.

Following the procedure outlined in Section 2.1, graphical ANOVAs can be constructed directly from a standard ANOVA table. Since it has not been common practice in our laboratory up to this point to add ANOVA tables explicitly to technical reports, we had to reconstruct the analyses documented by Hunter et al. (1995). The observant reader will notice the results of our analysis of variance (Table 5) differ from those reported by Hunter, et al. While it is unfortunate that we were not able to reproduce the original analysis, our results differ only slightly from those originally reported, and achieve the same patterns of significance.

Table 5: ANOVA table for JAERI II trial completion time.

Source	df	SS	MS	F	p
INT	1	3571846	3571846	8.93	0.0073
TRAIN	1	100423	100423	0.25	0.6217
INT*TRAIN	1	492897	492897	1.23	0.2801
SUBJECT(INT*TRAIN)	20	7997133	399856		
TRIAL	49	6341912	129426	6.12	0.0001
INT*TRIAL	49	863436	17621	0.83	0.7862
TRAIN*TRIAL	45	3765018	83667	3.96	0.0001
INT*TRAIN*TRIAL	44	938635	21332	1.01	0.4582
SUBJECT*TRIAL(INT*TRAIN)	897	18973025	21151		

To construct a graphical ANOVA first for the INT effect, recall that CIs are constructed as:

$$CI = M_j \pm \left(\sqrt{MS_\varepsilon / n} \right) t_{\alpha/2,\varphi} .$$

So, to construct a CI for the INT, the only unknowns left are the group means, which can be easily calculated from the data as 650.72 s for the P group and 533.47 s for the P+F group. For both groups, we use the mean square for the error term in the *F*-test, MS[SUBJECT(INT* TRAIN)], as MS_ε. Given that 12 subjects performed 50 trials[5], the *n* for this analysis is (12)(50) = 600. Finally, φ is equal to the error *df*, which is 20 in this case.

So, for the P group,

$$CI = 650.72 \pm \left(\sqrt{\frac{399856}{600}} \right) 2.086$$

$$= 650.72 \pm 53.85 ,$$

and for the P+F group,

$$CI = 533.47 \pm 53.85 .$$

Stated in more formal terms, we have:

$$\{596.87 \le \mu_P \le 704.57\}$$

$$\{479.62 \le \mu_{P+F} \le 587.32\} .$$

For the training effect, the means are 593.14 s for the no training group and 589.38 s for the AH training group. Since hypotheses on training are based on the same error term as

[5] While this is an unbalanced ANOVA, the TRIAL *df* still gives a good indication of what the number of trials was — less one — performed by all subjects.

hypotheses on interface, CIs on training will have the same width as CIs on interface (53.85 s). So,

$$\{539.29 \le \mu_{NO} \le 646.99\}$$

$$\{535.53 \le \mu_{AH} \le 643.23\}.$$

These confidence intervals have been plotted in Figures 5 and 6.

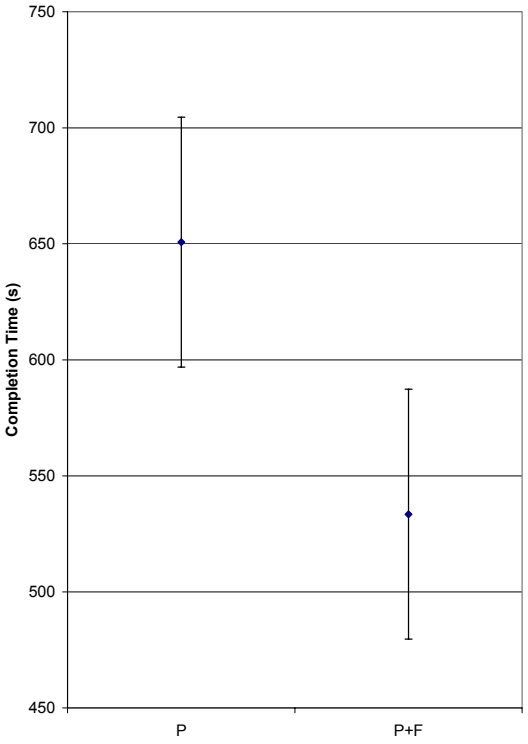

Figure 5: Graphical ANOVA on interface effect for JAERI II completion time.

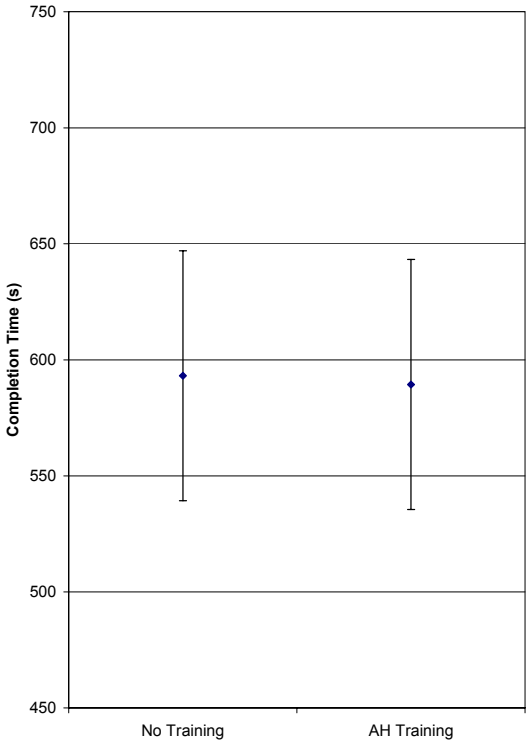

Figure 6: Graphical ANOVA on training effect for JAERI II completion time.

From these two charts, the four questions posed above can now be answered.

- **In general, how powerful is the conclusion on the INT effect?** Since there is a noticeable distance between the two CIs for the INT effect, it is obvious that this experiment had sufficient power to enable us to make conclusions from the data. However, since the distance between the CIs is, relatively speaking, quite small, obvious improvements can be made in experimental power. In fact, the power of this conclusion is about 0.77, just slightly below the value of 0.8 which is generally considered to be 'acceptable' power.

- **How close to significant (i.e., $p < .05$) was the training (TRAIN) effect?** Since the means of the two TRAIN conditions are very close together with respect to the width of the CIs, the *p*-value of 0.62 obtained for the TRAIN effect gains some meaning. While these two graphs do not output specific *p*-values, they do afford a much more intuitive interpretation of significance.
- **What was the pattern of observed means for the two effects?** While this relatively simple question finds no answer in the ANOVA table, it is quite obvious from these two graphs. Subject using the P interface were slower than those using the P+F interface, and subjects who underwent no training were only slightly slower than those who were given AH training.
- **What are the effect sizes for both INT and TRAIN?** Recall that effect size is the difference between the two group means divided by the variance in the data, or $|d/s|_{obs}$. While the information necessary to calculate this parameter is contained in the ANOVA table, the calculations are hardly intuitive. Notice, however, that both the distance between the means and the variance in the data is presented graphically on the CI plots, affording an intuitive calculation of effect size. From the two graphs above, it can be seen that the effect size for INT is at least moderate (i.e., greater than 0.4) while the effect size for the TRAIN effect is most certainly small (i.e., much less than 0.4). In fact, the effect size for INT is 0.44 and the effect size for TRAIN if 0.07, figures that are at least in line with the estimates given.

It could be argued that it is not practical to discard standard ANOVA reporting given the current standards of most journals in the behavioural sciences. While this might be the case, the example above has shown that confidence intervals at the very least complement standard ANOVA reports by providing intuitive information on some important parameters (such as power and effect size) that can only be extracted from the ANOVA table with some effort and by longhand calculation. CI plots also present information about the patterns in group means, data that simply cannot be found in an ANOVA table.

Power Analysis

Since power analyses have never before been performed on the data from the experiments performed for JAERI in previous contracts (Christoffersen et al., 1994; Howie et al., 1996;

Hunter et al., 1995), in this section the results of power analyses performed on selected tests for all of these experiments will be presented.

Power analyses were carried out for each of the JAERI experiments on inferences involving each of the six dependent measures introduced in the background. The results of these analyses will be broken down by experiment and then by dependent measure.

As was mentioned above, while the various experimenters were quite conscientious to report *F* and *p*-values for the analyses that turned out to be significant, useful records of the full ANOVA tables were not kept. As a result, we had to rerun the analyses. Our new analyses closely, though not exactly, match those reported previously.

JAERI I. In the JAERI I investigation, three effects are relevant to an analysis of power: the interface effect (INT), the effect of experience as coded by trial number (TRIAL), and the interaction between interface and experience (INT*TRIAL). A power analysis for TCT is reproduced in Table 6, along with values for ϕ and power (1-β). (To save space, the sums of squares have not been included in the ANOVA tables in this section, but these can easily be found by multiplying the mean squares by their corresponding degrees of freedom. In addition, to help in identifying those effects that have appreciable power, in the tables that follow all effects with a power of greater than 0.80 have been shaded.)

Table 6: Power analysis on TCT for JAERI I.

Source	df	MS	F	p	ϕ	1 - β
INT	1	131617	0.18	0.6907	undef.	< 0.30[6]
SUBJECT(INT)	4	696102				
TRIAL	204	79591	6.91	0.0001	2.43	> 0.99
INT*TRIAL	202	11773	1.03	1.03	0.17	< 0.30
TRIAL*SUBJECT(INT)	719	11454				

These results reveal that the TRIAL effect had a very high power, while the other effects achieved only low power. The high power of the trial effect indicates that the finding that subjects did learn over the course of the experiment is quite reliable.

Since longitudinal experiments like JAERI I are expensive and difficult to carry out, a more detailed power analysis on TCT was carried out. In this analysis, the experiment was split

[6] Power in this report is often reported as < 0.30. This is because most power charts do not give powers for low values of ϕ (e.g., Kirk, 1995; Pearson & Hartley, 1951)

Table 7: Block definitions

Trials	Block	Trials	Block
1-18	1	119-138	7
19-38	2	139-158	8
39-58	3	159-178	9
59-78	4	179-198	10
79-98	5	199-217	11
99-118	6		

Table 8: Power analysis on TCT by block for JAERI I.

Block	Effect	df_{effect}	df_{error}	MSW	MSE	p	ϕ	1-β
1	INT	1	4	108097	451408	0.65	undef.	< 0.30
1	TRIAL	17	52	152100	36385	< 0.01	1.73	0.94
1	INT*TRIAL	15	52	38150	36385	0.42	0.21	< 0.30
2	INT	1	4	39	31785	0.97	undef.	< 0.30
2	TRIAL	19	65	29848	7125	< 0.01	1.74	0.96
2	INT*TRIAL	19	65	4523	7125	0.86	undef.	< 0.30
3	INT	1	4	35982	57336	0.23	0.71	< 0.30
3	TRIAL	19	67	208153	12082	< 0.01	3.89	> 0.99
3	INT*TRIAL	19	67	17248	12082	0.14	0.64	< 0.30
4	INT	1	4	4742	48075	0.77	undef.	< 0.30
4	TRIAL	18	63	8504	8321	0.45	0.14	< 0.30
4	INT*TRIAL	18	63	10052	8321	0.28	0.44	< 0.30
5	INT	1	4	99371	20827	0.09	1.37	0.32
5	TRIAL	17	63	19660	10202	0.03	0.94	0.5
5	INT*TRIAL	17	63	1919	10202	0.99	undef.	< 0.30
6	INT	1	4	105285	38021	0.17	0.94	0.50
6	TRIAL	18	69	12757	6094	0.02	1.02	0.54
6	INT*TRIAL	18	69	7879	6094	0.22	0.53	< 0.30
7	INT	1	4	59989	66365	0.40	undef.	< 0.30
7	TRIAL	19	69	6045	5535	0.39	0.3	< 0.30
7	INT*TRIAL	19	69	4676	5535	0.65	undef.	< 0.30
8	INT	1	4	10554	44184	0.64	undef.	< 0.30
8	TRIAL	18	62	12802	8627	0.14	0.66	< 0.30
8	INT*TRIAL	18	62	13361	8627	0.10	0.72	< 0.30
9	INT	1	4	10846	78872	0.73	undef.	< 0.30
9	TRIAL	17	61	8288	5090	0.14	0.66	< 0.30
9	INT*TRIAL	17	61	4804	5090	0.53	undef.	< 0.30
10	INT	1	4	10	79631	0.99	undef.	< 0.30
10	TRIAL	16	53	9586	5841	0.09	0.78	< 0.30
10	INT*TRIAL	16	53	7204	5841	0.28	0.47	< 0.30
11	INT	1	4	2705	116667	089	undef.	< 0.30
11	TRIAL	16	55	7773	8437	0.55	undef.	< 0.30
11	INT*TRIAL	16	55	5311	8437	0.85	undef.	< 0.30

into 11 blocks of approximately 19 trials as shown in Table 7. The purpose of this analysis was to determine the number of blocks for which the TRIAL effect was powerful, as those are the

blocks in which a rapid rate of learning was most likely occurring. The results of this analysis are shown in Table 8.

From this table it can be seen that the only effect to achieve an appreciable power was TRIAL, and that this only occurred in blocks 1-3 (trials 1-58). This indicates that reasonable conclusions about experience effects can be made over the first three blocks only, or in other words, that dramatic learning effects can be seen in blocks 1-3 but not later. This is not to say that no learning occurs over blocks 4 and 11, but rather that the learning curve has flattened out in these blocks. This finding shows that the use of only 67 trials in JAERI II, IIIa, and IIIb is valid as a more efficient experimental strategy than the 217 trials of JAERI I. According to the above findings, the strategy allows enough time for subjects to achieve relatively stable performance.

Table 9: Power analysis on JAERI I CTV.

Source	df	MS	F	p	ϕ	1 - β
INT	1	23768	5.60	0.07	1.52	0.34
SUBJECT(INT)	4	4243				
BLOCK	10	18658	4.40	0.08	2.27	> 0.99
INT*BLOCK	10	2521	0.90	0.54	undef.	
BLOCK*SUBJECT(INT)	40	2807				

A power analysis was also performed on CTV (Table 9). The only effect to achieve appreciable power is BLOCK, the blocks of trials over which variance was calculated (see Table 7). Recall that one clear finding of the JAERI I investigation was that subjects using the P+F interface exhibited lower CTV than subjects using the P interface. These findings show that the power for that conclusion is quite low. As there were only three subjects per group in this experiment, this is not surprising. When the data from the ANOVA table was used to predict the power of this conclusion for greater numbers of subjects (see Kirk, 1995), power increased to 0.80 at 8 subjects per group.

Table 10: Power analysis on JAERI I DET.

Source	df	MS	F	p	ϕ	1 - β
INT	1	61485	2.25	0.21	0.79	< 0.30
SUBJECT(INT)	4	27367				
TRIAL	1	34252	1.30	0.32	0.39	< 0.30
TRIAL*INT	1	57417	2.18	0.21	0.77	< 0.30
TRIAL*SUBJECT(INT)	4	26309				

Table 11: Power analysis on JAERI I DA.

Source	df	MS	F	p	ϕ	1 - β
INT	1	4.23	11.39	0.03	2.28	0.72
SUBJECT(INT)	4	0.37				
TRIAL	8	0.90	0.91	0.53	undef.	< 0.30
TRIAL*INT	8	0.51	0.51	0.83	undef.	<0.30
TRIAL*SUBJECT(INT)	23	1.00				

Table 12: Power analysis on JAERI I DGT.

Source	df	MS	F	p	ϕ	1 - β
INT	1	1277	0.47	0.56	undef.	< 0.30
SUBJECT(INT)	4	2706				
TRIAL	1	3645	0.83	0.46	undef.	< 0.30
TRIAL*SUBJECT(INT)	3	3198				

Table 13: Power analysis on JAERI I CT.

Source	df	MS	F	p	ϕ	1 - β
INT	1	51439	2.66	0.18	0.91	< 0.30
SUBJECT(INT)	4	19361				
TRIAL	1	39125	3.39	0.14	1.09	< 0.30
TRIAL*INT	1	29095	2.52	0.19	0.89	< 0.30
SUBJECT(INT*TRIAL)	4	11524				

Power analyses were also performed on fault detection time (Table 10), diagnosis accuracy (Table 11), diagnosis time (Table 12), and compensation time (Table 13) for JAERI I. The power for all effects except INT for DA was low. Although the conclusion for DA only has a power of 0.72, considering that there were only three subjects per group this conclusion has a relatively high power. In fact, if only one subject were to be added per group, this conclusion would have a power of 0.87. This is an important result as one of the major findings of the JAERI I investigation was that the P+F interface supports fault diagnosis better than the P interface. That this conclusion has high power is another confirmation of its validity.

JAERI II. The JAERI II experiment included both INT and TRIAL effects, but also had a TRAIN effect with two levels (no training and AH training). In this investigation, subjects were nested within both INT and TRAIN, making for a slightly more complex experimental design than JAERI I.

A power analysis was first performed on TCT (Table 14).

Table 14: Power analysis on TCT for JAERI II.

Source	df	MS	F	p	ϕ	1 - β
INT	1	3571846	8.93	< 0.01	1.99	0.78
TRAIN	1	100423	0.25	0.62	undef.	< 0.30
INT*TRAIN	1	492897	1.23	0.28	0.34	< 0.30
SUBJECT(INT*TRAIN)	20	399856				
TRIAL	49	129426	6.12	< 0.01	2.24	> 0.99
INT*TRIAL	49	17621	0.83	0.79	undef.	< 0.30
TRAIN*TRIAL	45	83667	3.96	< 0.01	1.70	0.97
INT*TRAIN*TRIAL	44	21332	1.01	0.46	0.09	< 0.30
SUBJECT*TRIAL(INT*TRAIN)	897	21151				

Four things can be noticed from this analysis. First, the INT effect has nearly reached the benchmark power of 0.8 (adding one more subject per group would bring the power for this effect up to 0.82). This is a notable increase in power for the INT effect on this measure from JAERI I, where the same effect had a power of < 0.30. This increase can be attributed to the quadrupled sample size of JAERI II. Second, this experiment unfortunately had low power to make conclusions about the TRAIN effect. Third, the power for TRIAL was very high, indicating that this experiment was sensitive enough to detect experience effects. Finally, and to supplement the second result, this experiment had high power to make conclusions about the interaction between INT and TRAIN. This supports the conclusions made by Hunter, et al. (1995) on this interaction.

A power analysis on CTV is presented in Table 15.

Table 15: Power analysis on CTV for JAERI II.

Source	df	MS	F	p	ϕ	1-β
INT	1	35977	12.22	< 0.01	2.37	0.88
TRAIN	1	375	0.13	0.72	undef.	< 0.30
INT*TRAIN	1	217	0.07	0.79	undef.	< 0.30
SUBJECT(INT*TRAIN)	20	2943				

This table reveals that the INT effect had a much higher power than was observed in the JAERI I investigation. Again, this is most probably due to the quadrupling of the sample size in this experiment.

Power analyses on the four fault measures are shown in Tables 16-19.

Table 16: Power analysis on DET for JAERI II.

Source	df	MS	F	p	ϕ	1-β
INT	1	19221	0.77	0.39	undef.	< 0.30
TRAIN	1	18928	0.76	0.40	undef.	< 0.30
INT*TRAIN	1	248	0.01	0.92	undef.	< 0.30
SUBJECT(INT*TRAIN)	20	25051				
TRIAL	8	70052	6.92	< 0.01	2.28	> 0.99
INT*TRIAL	7	14950	1.48	0.19	0.64	< 0.30
TRAIN*TRIAL	8	17444	1.72	0.11	0.80	< 0.30
INT*TRAIN*TRIAL	7	8035	0.79	0.58	undef.	< 0.30
SUBJECT*TRIAL(INT*TRAIN)	100	10119				

Table 17: Power analysis on DA for JAERI II.

Source	df	MS	F	p	ϕ	1-β
INT	1	38.13	15.01	> 0.01	2.65	0.94
TRAIN	1	0.83	0.33	0.57	undef.	< 0.30
INT*TRAIN	1	4.73	1.86	0.19	0.66	< 0.30
SUBJECT(INT*TRAIN)	20	50.80				
TRIAL	8	33.19	6.28	> 0.01	2.15	> 0.99
INT*TRIAL	7	6.98	1.54	0.17	0.68	< 0.30
TRAIN*TRIAL	8	6.09	1.15	0.34	0.36	< 0.30
INT*TRAIN*TRIAL	7	4.54	1.00	0.43	0.05	< 0.30
SUBJECT*TRIAL(INT*TRAIN)	132	99.66				

Table 18: Power analysis on DGT for JAERI II.

Source	df	MS	F	p	ϕ	1-β
INT	1	103251	3.37	0.08	1.09	< 0.30
TRAIN	1	108110	3.53	0.08	1.13	< 0.30
INT*TRAIN	1	175541	5.74	0.03	1.54	0.50
SUBJECT(INT*TRAIN)	17	30593				
TRIAL	8	33169	1.87	0.10	0.87	< 0.30
INT*TRIAL	6	51604	2.91	0.03	1.26	0.76
TRAIN*TRIAL	8	21237	1.20	0.33	0.42	< 0.30
INT*TRAIN*TRIAL	4	33966	1.91	0.17	0.68	< 0.30
SUBJECT*TRIAL(INT*TRAIN)	41	17742				

Table 19: Power analysis on CT for JAERI II.

Source	df	MS	F	p	ϕ	1-β
INT	1	481209	3.17	0.09	1.04	< 0.30
TRAIN	1	511405	3.37	0.08	1.09	< 0.30
INT*TRAIN	1	17422	0.11	0.74	undef.	< 0.30
SUBJECT(INT*TRAIN)	20	151715				
TRIAL	8	766685	11.42	< 0.01	3.02	> 0.99
INT*TRIAL	7	9012	0.13	0.99	undef.	< 0.30
TRAIN*TRIAL	8	62128	0.93	0.49	undef.	< 0.30
INT*TRAIN*TRIAL	7	44517	0.66	0.68	undef.	< 0.30
SUBJECT*TRIAL(INT*TRAIN)	132	67142				

Fault trials in this investigation achieved higher power than those of JAERI I, most likely because of the increase in sample size between the two experiments. The strongest finding here is that for three of the measures (DET, DA, and CT) the TRIAL effect has a very high power (> 0.99 in all three cases). The quadrupling of sample size has opened up the opportunity of making reliable claims on the effect of learning in fault performance. The only other measure/effect combination to achieve a high power was the INT effect on diagnosis accuracy. Again, that this conclusion has high power is a validation of the conclusion that the P+F interface induces more effective fault diagnosis than the P interface.

JAERI II as a Mini-Experiment. A second power analysis on the data from JAERI II was performed, this time considering only the data from the no-training group. Isolating this group leaves a dataset similar to the JAERI I investigation, but with double the sample size. This analysis should move us one step further towards understanding what sample size is appropriate for investigations with DURESS II.

For this reduced data set, power analyses were performed on all six measures, and are documented in Tables 20-25.

Table 20: Power analysis on TCT for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	5357889	11.46	< 0.01	2.51	.89
SUBJECT(INT)	10	467390				
TRIAL	61	298425	6.02	< 0.01	2.22	> 0.99
TRIAL*INT	61	68816	1.39	0.03	0.62	< 0.30
SUBJECT(INT*TRIAL)	574	49557				

Table 21: Power analysis on CTV for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	15300	4.38	0.06	1.30	0.38
SUBJECT(INT)	10	3493				

Table 22: Power analysis on DET for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	6188	0.39	0.55	undef.	< 0.30
SUBJECT(INT)	10	158580				
TRIAL	7	19729	2.64	0.02	1.20	0.76
TRIAL*INT	6	1618	0.22	0.97	undef.	< 0.30
TRIAL*SUBJECT(INT)	44	7468				

Table 23: Power analysis on DA for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	34.00	19.33	< 0.01	3.03	0.96
SUBJECT(INT)	10	1.76				
TRIAL	7	4.32	6.85	< 0.01	2.94	> 0.99
TRIAL*INT	6	1.68	2.66	0.02	1.19	0.75
TRIAL*SUBJECT(INT)	59	0.63				

Table 24: Power analysis on DGT for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	1124	0.03	0.86	undef.	< 0.30
SUBJECT(INT)	8	36419				
TRIAL	7	22390	1.98	0.11	0.92	< 0.30
TRIAL*INT	4	42514	3.75	0.02	1.55	0.83
TRIAL*SUBJECT(INT)	21	11335				

Table 25: Power analysis on CT for JAERI II no training data.

Source	df	MS	F	p	ϕ	1 - β
INT	1	337100	3.09	0.11	1.02	< 0.30
SUBJECT(INT)	10	108936				
TRIAL	7	310756	4.59	< 0.01	1.77	0.96
TRIAL*INT	6	37563	0.55	0.76	undef.	< 0.30
TRIAL*SUBJECT(INT)	57	67776				

Not surprisingly, increases in power can be observed almost across the board when these results are compared to those from JAERI I. These data achieve appreciable power for: (1) the INT and TRIAL effects for TCT, (2) the INT and TRIAL effects for DA, and (3) the TRIAL effect for CT. The TRIAL effect for DET (power of 0.76) is predicted to reach appreciable power at 7 subjects per group.

These data also achieved appreciable power for the TRIAL*INT interaction for DGT. Unfortunately, hypotheses could not be tested on this effect for JAERI I due to a lack of degrees of freedom. The only measure for which the JAERI I investigation was more powerful was CTV. This is understandable as the JAERI I investigation had many more trials over which to calculate variance.

JAERI IIIa. This experiment used a similar experimental design to JAERI I, with the INT effect now representing the P+F and divided P+F interfaces. Power analyses were performed on the data from this experiment (Tables 26 - 31).

Table 26: Power analysis on TCT for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	5144093	10.74	< 0.01	2.21	0.81
SUBJECT(INT)	10	478783				
TRIAL	55	421850	12.50	< 0.01	3.36	> 0.99
TRIAL*INT	55	34799	1.03	0.41	0.17	< 0.30
TRIAL*SUBJECT(INT)	128	33756				

Table 27: Power analysis on CTV for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	16937	3.36	0.09	1.09	< 0.30
SUBJECT(INT)	10	5036				

Table 28: Power analysis on DET for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	2378	0.56	0.47	undef.	< 0.30
SUBJECT(INT)	10	4245				
TRIAL	8	47328	26.22	< 0.01	4.73	> 0.99
TRIAL*INT	8	4122	2.28	0.04	1.07	0.50
TRIAL*SUBJECT(INT)	49	1805				

Table 29: Power analysis on DA for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	0.039	0.01	0.91	undef.	< 0.30
SUBJECT(INT)	10	3.072				
TRIAL	8	2.863	2.93	< 0.01	1.31	0.75
TRIAL*INT	8	1.516	1.55	0.15	0.70	< 0.30
TRIAL*SUBJECT(INT)	80	0.976				

Table 30: Power analysis on DGT for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	132	0.01	0.93	undef.	< 0.30
SUBJECT(INT)	8	17209				
TRIAL	8	26264	1.03	0.48	0.16	< 0.30
TRIAL*INT	5	35927	1.41	0.32	0.60	< 0.30
TRIAL*SUBJECT(INT)	8	25548				

Table 31: Power analysis on CT for JAERI IIIa.

Source	df	MS	F	p	ϕ	1 - β
INT	1	1059499	10.78	< 0.01	2.21	0.81
SUBJECT(INT)	10	98303				
TRIAL	5	613984	17.99	< 0.01	3.76	> 0.99
TRIAL*INT	5	100271	2094	0.02	1.27	0.65
TRIAL*SUBJECT(INT)	49	34120				

These results show that appreciable power was achieved for the measures of TCT and CT for both INT and TRIAL effects, and for DET on only the TRIAL effect. In comparing these results to those of JAERI I and the JAERI II mini-experiment, it is interesting to note that this investigation achieved higher power on time-based measures (TCT and CT) while JAERI I and the JAERI II mini experiment achieved higher power on fault-based measures (DET, but especially DA and DGT). This is most easily explained by contrasting the different experimental manipulations of these investigations. In JAERI I and the JAERI II mini experiment, a traditional interface (P) was being compared against one specially designed to support operators in abnormal situations (P+F), while in JAERI IIIa all subjects had access to the informational enhancements of the P+F interface except that subjects in one group used an interface that split up the levels of information. As a result, a slight time penalty was incurred for the one group when switching between these levels. Thus, it is not surprising to see that JAERI I and the JAERI II mini-experiment had power to detect differences in fault performance while the JAERI IIIa experiment had power to detect differences in trial completion time either in normal or fault conditions.

JAERI IIIb. The last power analyses were performed on the data from the JAERI IIIb investigation. The experimental design here is similar to JAERI IIIa, except that the main effect is now TRAIN and represents the three training groups of this experiment (no training, replay, and replay plus self-explanation). The results of these analyses are shown in Tables 32 - 37.

Table 32: Power analysis on TCT for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	56164	0.05	0.96	undef.	< 0.30
SUBJECT(TRAIN)	15	1233021				
TRIAL	55	336458	12.10	< 0.01	3.30	> 0.99
TRIAL*TRAIN	109	25779	0.93	0.67	undef.	< 0.30
TRIAL*SUBJECT(TRAIN)	779	27808				

Table 33: Power analysis on CTV for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	2349	0.32	0.73	undef.	< 0.30
SUBJECT(TRAIN)	15	7231				

Table 34: Power analysis on DET for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	14789	1.03	0.38	0.14	< 0.30
SUBJECT(TRAIN)	14	14390				
TRIAL	8	37281	4.23	< 0.01	1.71	0.93
TRIAL*TRAIN	16	11220	1.27	0.24	0.51	< 0.30
TRIAL*SUBJECT(TRAIN)	74	8819				

Table 35: Power analysis on DA for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	5.79	1.48	0.26	0.56	< 0.30
SUBJECT(TRAIN)	15	3.92				
TRIAL	8	1.87	1.71	0.10	0.70	< 0.30
TRIAL*TRAIN	16	1.38	1.26	0.23	0.42	< 0.30
TRIAL*SUBJECT(TRAIN)	120	1.09				

Table 36: Power analysis on DGT for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	72849	6.33	0.01	1.88	0.74
SUBJECT(TRAIN)	13	11517				
TRIAL	8	38219	1.21	0.34	0.44	< 0.30
TRIAL*TRAIN	15	43724	1.39	0.24	0.60	< 0.30
TRIAL*SUBJECT(TRAIN)	25	31500				

Table 37: Power analysis on CT for JAERI IIIb.

Source	df	MS	F	p	ϕ	1 - β
TRAIN	2	35300	0.16	0.85	undef.	< 0.30
SUBJECT(TRAIN)	15	219587				
TRIAL	6	355754	9.33	< 0.01	2.67	> 0.99
TRIAL*TRAIN	12	16482	0.43	0.93	undef.	< 0.30
TRIAL*SUBJECT(TRAIN)	86	38124				

From these results we can see that in terms of detecting differences in trial completion time and fault detection and diagnosis, JAERI IIIb was a weak experiment. For the measures of TCT, DET, and CT, TRIAL is the only effect to achieve appreciable power. These results indicate that conclusions about learning effects can reasonably be made on time-based measures. It should be noted that this experiment was relatively powerful as well in terms of the INT effect for DGT, which is predicted to only require one more subject to reach a power of 0.80.

Discussion. The results of these power analyses are encouraging. Across the four experiments for which analyses were performed, many of the conclusions made by previous

experimenters have been shown to have appreciable power. In the JAERI I, and especially in the JAERI II mini experiment, fault-based measures achieved high power, validating the conclusions made on the extra support for abnormal situations afforded by the P+F interface over the P interface. Further, in the JAERI II mini-experiment, the INT effect on CTV achieved appreciable power, and in JAERI I the same effect/measure combination had a relatively high power considering the small sample size. These results validated the conclusions made to the effect that the P+F interface induces more consistent behaviour. It is reassuring to have revealed this reconfirmation of these results.

As these analyses involved only time- and fault-based measures, few of the major conclusions from JAERI II, JAERI IIIa, or JAERI IIIb were validated. While this is unfortunate, for future experiments it is useful understand that time- and fault-based measures might not be sensitive to changes induced by training and information form manipulations like those found in these experiments.

These analyses have also made an important methodological contribution to this research programme. First, the power analyses by block for JAERI I confirmed that 67 trials (that is, about a month of experimentation time) is a suitable duration for experiments. While experience from the JAERI I investigation has demonstrated that subjects still learn new strategies well past trial 67, the analyses introduced here support 67 trials as a more efficient experimental duration than the 217 trials of JAERI I. Second, these analyses provide a motivation for a small increase in sample size. In four instances (the INT effect for JAERI I CTV and JAERI II TCT, the TRIAL effect for the JAERI II mini experiment, and the TRAIN effect for JAERI IIIb DGT) power can be predicted to increase to > 0.80 with the addition of one or two subjects per group. Thus, these analyses indicate that future experiments should be formed from groups of 7-8 subjects, each performing 67 trials on DURESS II.

Bayesian Methods to Assert Largeness or Smallness

The final set of analyses performed during this study used the Bayesian methods of Rouanet (1996) to determine and make inferences on the effect sizes for the six measures described above. Since we were not able to obtain the software needed to perform multi-*df* analyses, we were limited to working with the data for 1-*df* effects only. Fortunately, 1-*df* effects are fairly common in the research that we have performed for JAERI, and this restriction still

allowed us to analyse the main effects from all experiments except JAERI IIIb (which had a 2-*df* main effect). The results of these analyses are presented below in three formats.

Table 38 first presents a chart containing the effect sizes and the data from which they were derived. Though not terribly informative, this table is included to help the reader in reconstructing any of the analyses, should she want to. Figures 7 – 11 present the results from the chart as graphs, broken down by experiment. Finally, Figures 12 – 17 present the same results broken down by measure, across experiments. A discussion of the results follows these tables and figures.

Table 38: 1-*df* effect size analyses.

Expt	Meas	Effect	n	MS_{effect}	MS_{error}	Subj. / Group	df_{error}	$(d/s)_{obs}$	CI Upper	CI Lower
J1	TCT	INT	1131	131617.0	696102.0	3	4	0.032	0.112	0.032
	CTV	INT	1131	23768.0	4243.0	3	4	0.172	0.112	0.112
	DET	INT	52	61485.6	27368.0	3	4	0.509	0.521	0.509
	DA	INT	54	4.2	0.4	3	4	1.125	0.511	0.511
	DGT	INT	19	1277.0	2706.4	3	4	0.386	0.862	0.386
	CT	INT	54	51439.0	19361.0	3	4	0.543	0.511	0.511
J2	TCT	INT	1108	3571846.0	399856.0	12	20	0.440	0.195	0.195
	CTV	INT	1108	35977.0	2943.1	12	20	0.515	0.195	0.195
	DET	INT	189	19221.0	25051.0	12	20	0.312	0.472	0.312
	DA	INT	258	38.1	2.5	12	20	1.182	0.404	0.404
	DGT	INT	100	103251.0	30593.0	12	20	0.900	0.649	0.649
	CT	INT	186	481209.0	151715.0	12	20	0.640	0.476	0.476
	TCT	TRAIN	1108	100423.0	399856.0	12	20	0.074	0.195	0.074
	CTV	TRAIN	1108	376.0	2943.1	12	20	0.053	0.195	0.053
	DET	TRAIN	189	18928.0	25051.0	12	20	0.310	0.472	0.310
	DA	TRAIN	258	0.8	2.5	12	20	0.175	0.404	0.175
	DGT	TRAIN	100	108110.0	30593.0	12	20	0.921	0.649	0.649
	CT	TRAIN	186	511405.0	151715.0	12	20	0.660	0.476	0.476
J2-MINI	TCT	INT	708	5354889.4	467390.9	6	10	0.441	0.179	0.179
	CTV	INT	708	15300.0	3493.0	6	10	0.272	0.179	0.179
	DET	INT	97	6188.0	15858.0	6	10	0.220	0.483	0.220
	DA	INT	132	34.0	1.8	6	10	1.326	0.414	0.414
	DGT	INT	58	1124.0	36419.9	6	10	0.080	0.624	0.080
	CT	INT	94	337100.0	108936.0	6	10	0.629	0.490	0.490
J3a	TCT	INT	684	5144093.0	478783.0	6	10	0.434	0.182	0.182
	CTV	INT	684	16937.0	5036.0	6	10	0.243	0.182	0.182
	DET	INT	89	2378.0	4245.0	6	10	0.275	0.504	0.275
	DA	INT	144	0.0	3.1	6	10	0.032	0.396	0.032
	DGT	INT	39	132.7	17209.9	6	10	0.049	0.761	0.049
	CT	INT	83	1059499.3	98303.0	6	10	1.248	0.522	0.522

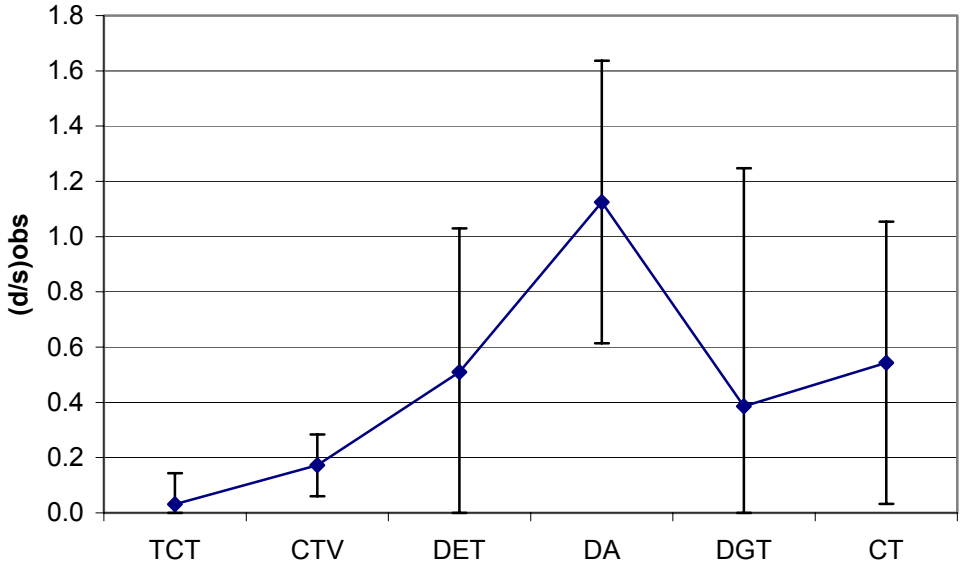

Figure 7: Effect size analysis on JAERI I INT effect.

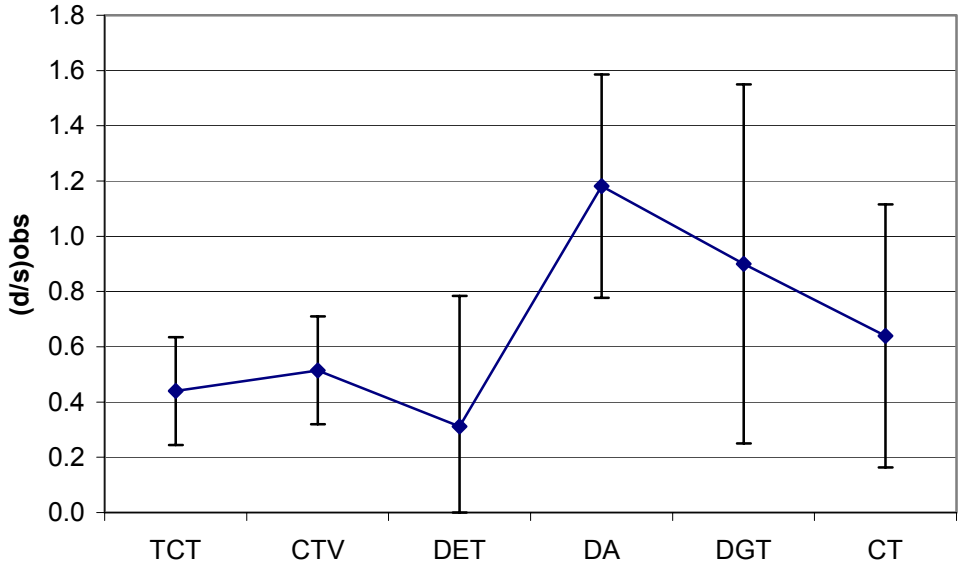

Figure 8: Effect size analysis on JAERI II INT effect.

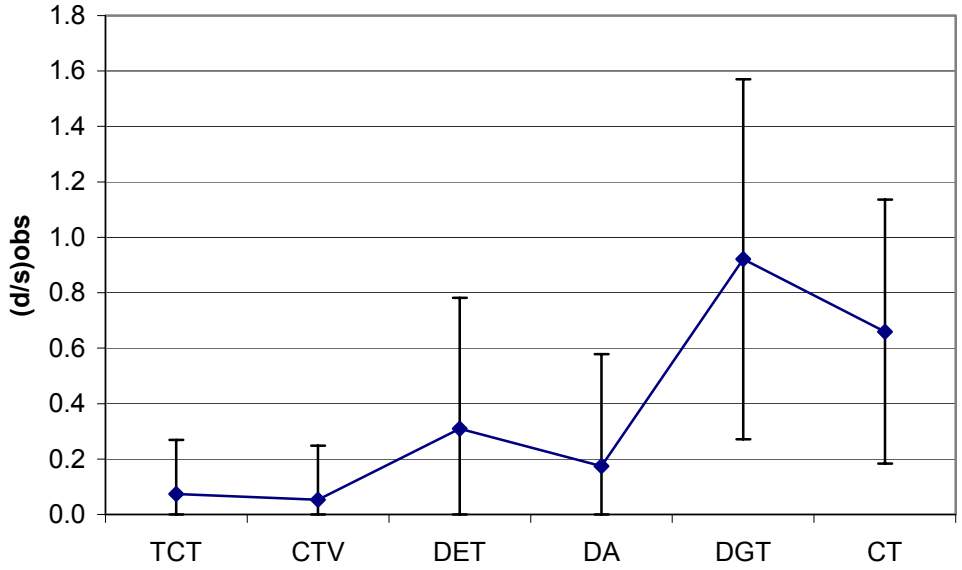

Figure 9: Effect size analysis on JAERI II TRAIN effect.

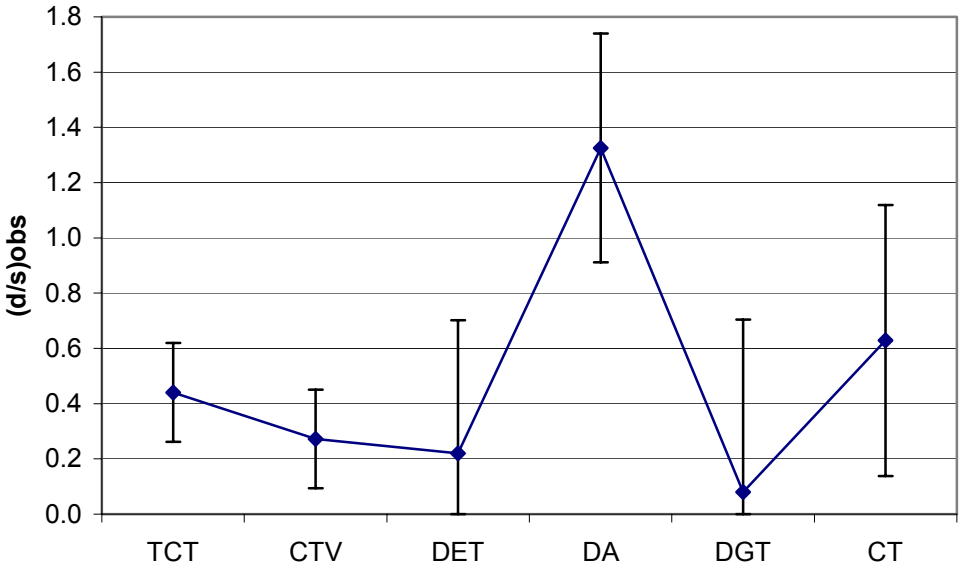

Figure 10: Effect size analysis on JAERI II mini experiment INT effect.

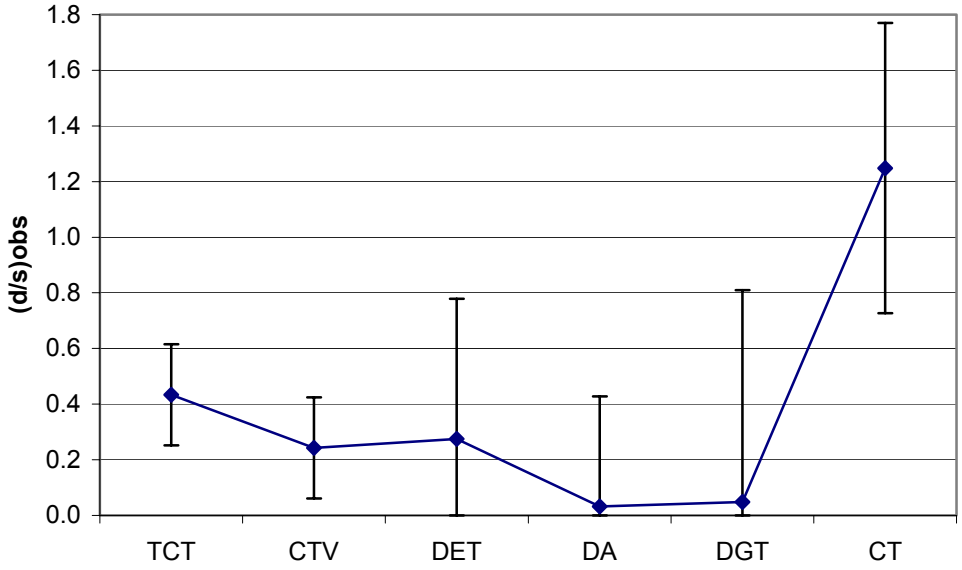

Figure 11: Effect size analysis on JAERI IIIa INT effect.

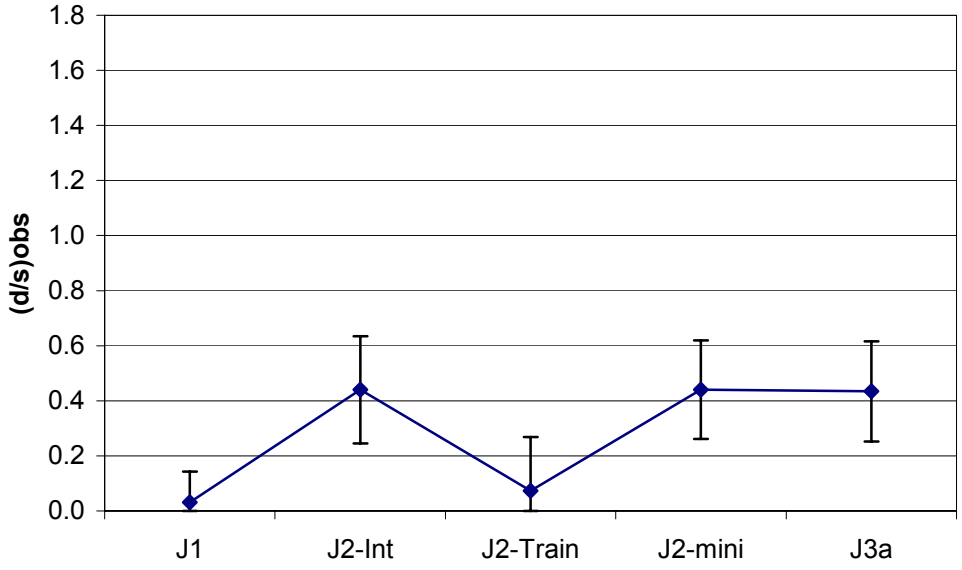

Figure 12: Effect size analysis on TCT across experiments.

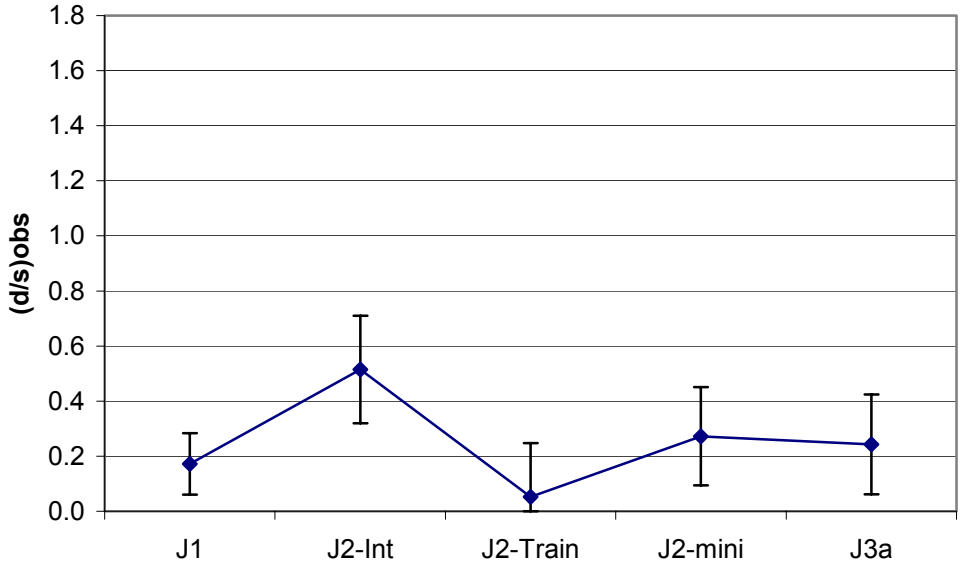

Figure 13: Effect size analysis on CTV across experiments.

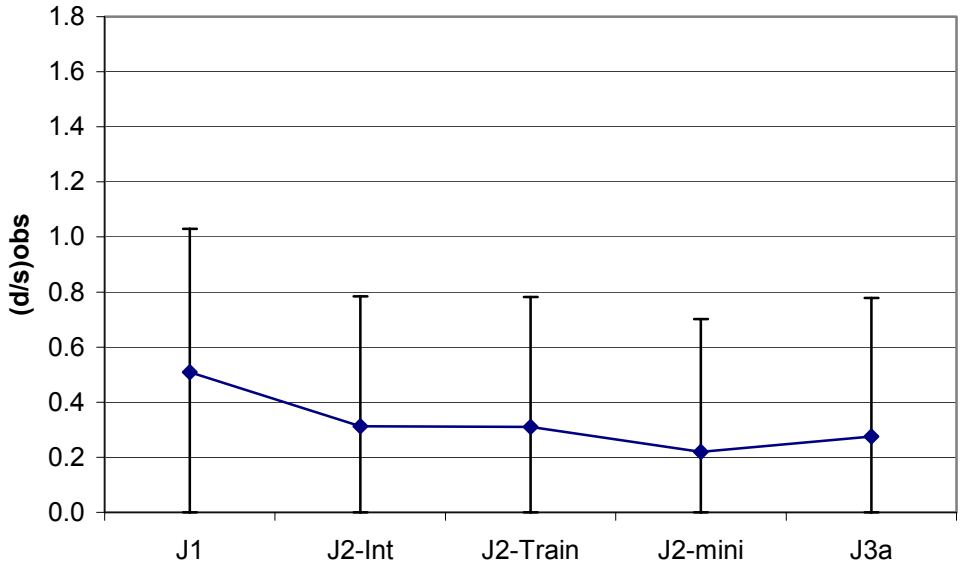

Figure 14: Effect size analysis on DET across experiments.

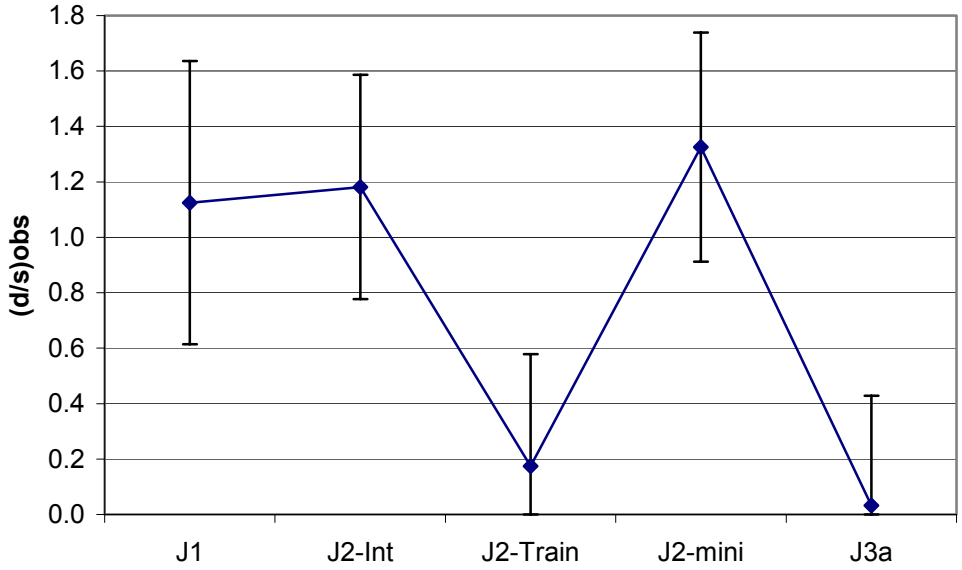

Figure 15: Effect size analysis on DA across experiments.

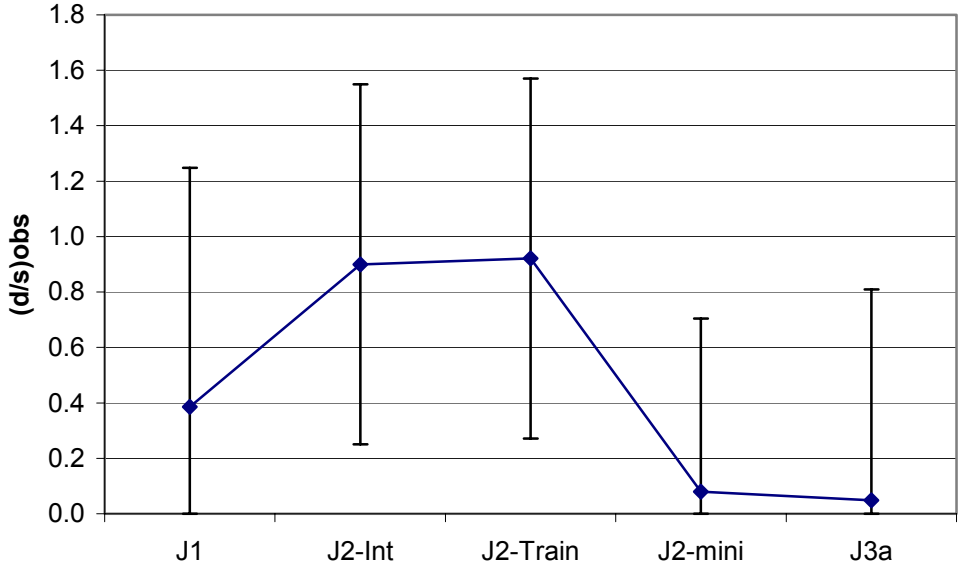

Figure 16: Effect size analysis on DGT across experiments.

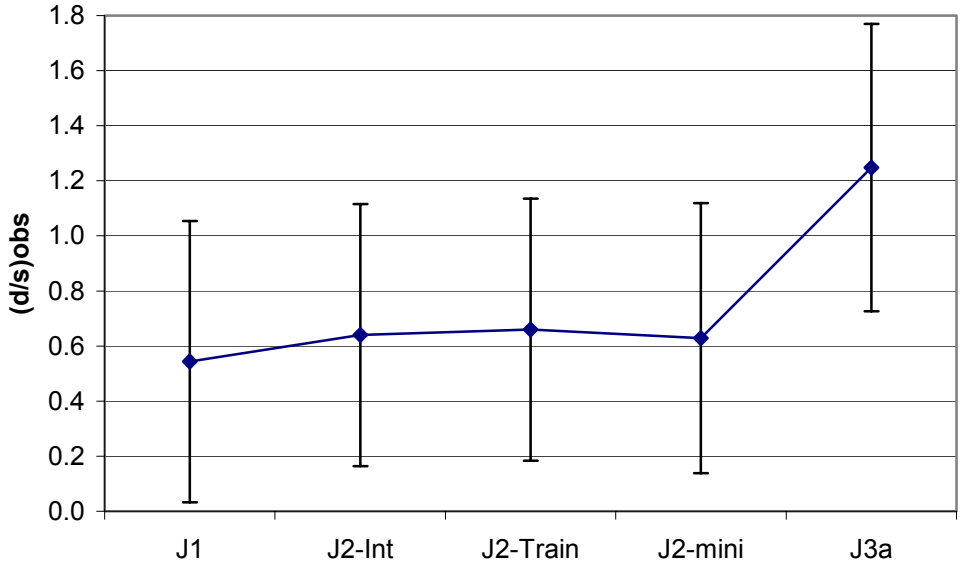

Figure 17: Effect size analysis on CT across experiments.

A number of interesting results can be found in these figures. We consider first Figures 7 – 11 which detail the effect sizes within experiments, and help to reveal the largest effects within each experiment. Four results stand out. First, in the experiments comparing the P interface against the P+F interface (Figures 7, 8, and 10) the interface effect on diagnosis accuracy is large. In other words, not only is the effect of the P+F interface on diagnosis accuracy reliable, it is also large. Second, in the JAERI II experiment, training had the largest effect on diagnosis time (Figure 9). Unfortunately, the graph clearly shows that we cannot assert largeness at the $\gamma = 0.9$ level. Third, in the JAERI IIIa experiment, the interface effect on compensation time was large (Figure 11). Although a portion of the size of this effect can be accounted for by the time cost incurred in switching between interface levels, this is not the whole story. Consider that the interface effect on TCT was small to moderate, and that this effect also includes the time cost of switching between levels. If this is the case, then some other factor must have come into play to make the effect of interface on compensation time so larger than its effect on trial completion time. While Howie, et al. (1996) do not address this point, it is likely that the divided interface had the effect of slowing subjects' ability to make decisions because the information was divided over four displays. Finally, it is notable that the effect of interface on completion time variance was not large in any of the experiments comparing the P and the P+F interfaces. This result was

not expected, but can perhaps be explained by the method used in this study to analyse variances. So that we could use the measures researched, an ANOVA was performed on completion time variance, when it should more properly be analysed using a Cochran test for the homogeneity of variance (Winer, 1962/1971). Since the test used was not well suited to the data in this case, the results are not necessarily valid.

A second set of results can be read from the graphs comparing effect sizes within measures and across experiments (Figures 12 - 17). Five results stand out here. First, the interface effect on trial completion time (Figure 12) was small for JAERI I and small to moderate for JAERI II and IIIa. The training effect on trial completion time for JAERI II was also small. Second, the largest effect size for completion time variance was observed in the JAERI II experiment for the interface effect. Third, little can be said about the effect sizes for fault detection time (Figure 14) and fault diagnosis time (Figure 16) as there a large amount of uncertainty in these data. In fact, larger uncertainty can be noticed for all of the measures for fault trials when compared to the measures for normal trials as there was a smaller dataset for these calculations (recall that Bayesian CIs widen as n_{eff} is reduced). Fourth, although there is much uncertainty in the effect sizes for fault compensation time, the effect of interface on this measure in JAERI IIIa was large (see above for an explanation). Fifth, and most importantly, Figure 15 demonstrates that experiments comparing the P interface against the P+F interface had large interface effects on diagnosis accuracy. Not only is the interface effect on diagnosis accuracy reliable for these experiments, it is also large.

One final analysis was performed to test one of the statements made by Rouanet (1996) about this method of testing effect sizes: "when the observed effect is very small, getting a statistically significant result — far from ruling out the conclusion of a small effect — is actually a cue for it" (p. 152). He calls this the *negligibility paradox*. As in this work we calculated a large pool of effect sizes that we have *p*-values to compare against (see

Table 38), we decided to test this assertion. This was done by plotting the *p*-values against their corresponding effect size. The result of this analysis is shown in Figure 18.

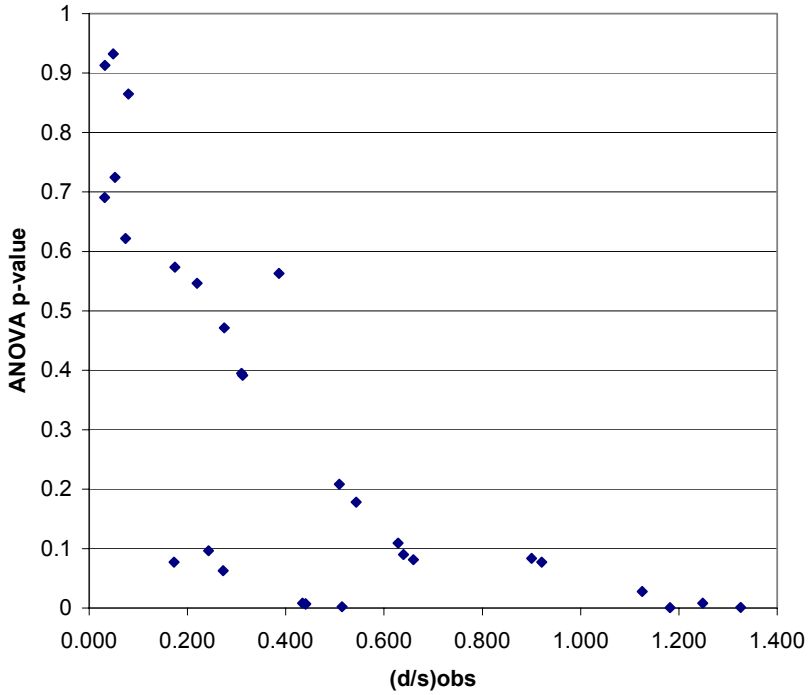

Figure 18: ANOVA *p*-values vs. effect sizes.

Figure 18 shows that our data do not at all confirm Rouanet's assertion. If the negligibility paradox were correct, there should be a positive correlation between *p*-values and effect sizes, while Figure 18 shows a negative correlation. A correlation analysis revealed a highly significant correlation ($p < 0.0001$) of –0.71 between *p*-values and effect sizes, indicating that our data show considerable deviation from this paradox.

Conclusions

In the preceding section, we have presented the results of applying three lesser-known data analysis techniques to our dataset. It is our belief that the addition of these techniques to our arsenal of data analysis techniques has been, and will continue to be, very beneficial. First of all, both the use of CIs and the calculation of effect sizes are powerful aids to data description. The examples given show how CI estimation helps the experimenter intuitively to consider the patterns of means, experimental power, and effect size without losing sight of simple inferences on statistical significance. The calculation of effect sizes has provided a new way to look at our data that can help in understanding which effects show promise of being large independent of statistical significance, thereby measuring practical significance.

Second, the power analyses conducted on much of the data from the JAERI experiments have confirmed that many of the conclusions of this research effort are powerful. Most notably, high power was achieved on a number of fault-based measures, confirming that the P+F interface does support operators in abnormal situations. Third, and related to this, the calculation of effect sizes has demonstrated that in addition to being powerful and significant, a number of the effects observed over the course of this program of research have been large. Most notable here are the large interface effects on diagnosis accuracy. Thus, the benefits of using the P+F interface, and the potential benefits of using EID in the design of interfaces for complex systems, can defensibly be said to be large.

Finally, the power analyses have also provided some direction for future experiments. First of all, there is a strong indication that subjects are able to obtain a fairly stable level of expertise in an experiment of 67 trials. While these shorter experiments do not provide the same opportunity for observing long-term adaptation as do longer experiments (like JAERI I), they are more manageable and allow the testing of a sufficient number of subjects to obtain good experimental power. Second, in order to obtain good power in a variety of measures, much will be gained from increasing group size from the current standard of 6 subjects to 7 or 8.

CONCLUSIONS

The purpose of the literature review presented in this review is not to point a finger at individuals who have relied on NHST or ANOVA to analyse data. We are just as guilty of uncritically using these techniques as anyone else. After all, these are the techniques that we have all been taught, that are well known by journal editors and reviewers, and that our statistics packages support. Thus, there are many pressures that cause people to continue to use the traditional methods. As Loftus (1991) put it, “The more you reject the null hypothesis, the more likely it is that you’ll get tenure” (p. 103).

After all, using some of the data analysis methods we have proposed here requires that we design our experiments differently too. If we are going to conduct an individual level analysis like Vicente (1992) and Hammond et al. (1987) did, then we need to rely more on within-subjects designs. If we are going to use the MTMM advocated by Campbell & Fiske (1959), we need to have multiple constructs and multiple methods in a single experiment. If we are going to be able to make point or interval predictions, we are going to have to develop stronger theories to guide experimentation. If we are going to develop a more cumulative knowledge base, we are

going to have to engage in more replication than we have in the past. Thus, a change in data analysis techniques is not a cosmetic change to be taken lightly. Instead, it requires some deep changes in the way in research is conducted.

Because of the enormity of this task, researchers typically find it easier to stick to what they are most comfortable with. 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 human factors engineers analyse their research data will not be easy: "Nothing short of a revolution will be required to escape from the methodological cul-de-sac into which the practice of hypothesis testing has led us" (Loftus & McLean, 1997, p. 151).

Nevertheless, the winds of change have begun. As editor of the journal Memory & Cognition, Loftus (1993a) has strongly encouraged authors to adopt non-traditional data analysis and presentation methods. Furthermore, the American Psychological Association has recently struck a committee to examine the issue of what methods are best suited to analysing behavioural data. Thus, after at least four decades, it seems that the revolution is finally gaining some potency.

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 & Rosenthal, 1989, p. 1282)

REFERENCES

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

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. The Psychological Review, 1, 27-53.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. The Psychological Review, 6, 347-375.

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

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

Christoffersen, K., Hunter, C. N., & 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 ed.). 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.

Crossman, E. R. F. W., & Cooke, J. E. (1962/1974). Manual control of slow-response systems. In E. Edwards & F. P. Lees (Eds.), The human operator and process control (pp. 51-66). London: Taylor and Francis.

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

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., & 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., & 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., & Steiger, J. H. (1997). What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum Associates.

Hines, W. W., & Montgomery, D. C. (1990). Probability and statistics in engineering and management science (Third ed.). Toronto: John Wiley & Sons.

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

Howie, D. E., Janzen, M. E., & Vicente, K. J. (1996). Research on factors influencing cognitive behaviour (III) (CEL 96-06). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Hunter, C. N., Janzen, M. E., & Vicente, K. J. (1995). Research on factors influencing human cognitive behaviour (II) (CEL 95-08). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences (Third ed.). Pacific Grove, CA.: Brooks/Cole Publishing Company.

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.

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. (in press). Psychology will be a much better science when we change the way we analyze data. Current Directions in Psychological Science (pp. 161-171).

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin & Review, 1, 476-490.

Loftus, G. R., & 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, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-197). Mahwah, NJ: Lawrence Erlbaum Associates.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition (pp. 1-53). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pawlak, W. S., & Vicente, K. J. (1996). Inducing effective operator control through ecological interface design. International Journal of Human-Computer Studies, 44, 653-688.

Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central *F*-distribution. Biometrika, 38, 112-130.

Phillips, L. D. (1973). Bayesian statistics for social scientists. London: Nelson.

Rosnow, R. L., & 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, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 175-197). Mahwah, NJ: Lawrence Erlbaum Associates.

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, 416-28.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 335-391). Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, F. L., & 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, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40, 73-83.

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

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 221-257). Mahwah, NJ: Lawrence Erlbaum Associates.

Venda, V. F., & 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. (in press). Four reasons why the science of psychology is still in trouble. Behavioral and Brain Sciences.

Winer, B. J. (1962/1971). Statistical principles in experimental design. New York: McGraw Hill.

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

Xiao, Y., & Vicente, K. J. (1997). Epistemological analysis in empirical (laboratory and field) studies: Data and generalizations. Manuscript submitted for publication.

Cognitive Engineering Laboratory

CEL TECHNICAL REPORT SERIES

CEL 93-01 "Egg-sucking, Mousetraps, and the Tower of Babel: Making Human Factors Guidance More Accessible to Designers"
• Kim J. Vicente, Catherine M. Burns, & William S. Pawlak

CEL 93-02 "Effects of Expertise on Reasoning Trajectories in an Abstraction Hierarchy: Fault Diagnosis in a Process Control System"
• Klaus Christoffersen, Alex Pereklita, & Kim J. Vicente

CEL 94-01 "Cognitive 'Dipsticks': Knowledge Elicitation Techniques for Cognitive Engineering Research"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-02 "Muddling Through Wicked Problems: Exploring the Role of Human Factors Information in Design"
• Catherine M. Burns

CEL 94-03 "Cognitive Work Analysis for the DURESS II System"
• Kim J. Vicente & William S. Pawlak

CEL 94-04 "Inducing Effective Control Strategies Through Ecological Interface Design"
• William S. Pawlak

CEL 94-05 "Research on Factors Influencing Human Cognitive Behaviour (I)"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-06 "Ecological Interfaces for Complex Industrial Plants"
• Nick Dinadis & Kim J. Vicente

CEL 94-07 "Evaluation of a Display Design Space: Transparent Layered User Interfaces"
• Beverly L. Harrison, Hiroshi Ishii, Kim J. Vicente, & Bill Buxton

CEL 94-08 "Designing and Evaluating Semi-Transparent 'Silk' User Interface Objects: Supporting Focused and Divided Attention"
• Beverly L. Harrison, Shumin Zhai, Kim J. Vicente, & Bill Buxton

CEL 95-01 "An Ecological Theory of Expertise Effects in Memory Recall"
• Kim J. Vicente & JoAnne H. Wang

CEL-SP "Strategic Plan"
• Cognitive Engineering Laboratory

CEL-LP "Cognitive Engineering Laboratory Profile"
• Cognitive Engineering Laboratory

CEL 95-04 "A Field Study of Operator Cognitive Monitoring at Pickering Nuclear Generating Station-B"
• Kim J. Vicente & Catherine M. Burns

CEL 95-05 "An Empirical Investigation of the Effects of Training and Interface Design on the Control of Complex Systems"
• Christopher N. Hunter

CEL 95-06 "Applying Human Factors to the Design of Medical Equipment: Patient-Controlled Analgesia"
• Laura Lin, Racquel Isla, Karine Doniz, Heather Harkness, Kim J. Vicente, & D. John Doyle

CEL 95-07 "An Experimental Evaluation of Transparent Menu Usage"
• Beverly L. Harrison & Kim J. Vicente

CEL 95-08 "Research on Factors Influencing Human Cognitive Behaviour (II)"
• Christopher N. Hunter, Michael E. Janzen, & Kim J. Vicente

CEL 95-09 "To the Beat of a Different Drummer: The Role of Individual Differences in Ecological Interface Design"
• Dianne Howie

CEL 95-10 "Emergent Features and Temporal Information: Shall the Twain Ever Meet?"
• JoAnne H. Wang

CEL 95-11 "Physical and Functional Displays in Process Supervision and Control"
• Catherine M. Burns & Kim J. Vicente

CEL 96-01 "Shaping Expertise Through Ecological Interface Design: Strategies, Metacognition, and Individual Differences"
• Dianne E. Howie

CEL 96-02 "Skill, Participation, and Competence: Implications of Ecological Interface Design for Working Life"
• Peter Benda, Giuseppe Cioffi, & Kim J. Vicente

CEL 96-03 "Practical Problem Solving in a Design Microworld: An Exploratory Study'
• Klaus Christoffersen

CEL 96-04 "Review of Alarm Systems for Nuclear Power Plants"
• Kim J. Vicente

CEL 96-05 "DURESS II User's Manual: A Thermal-hydraulic Process Simulator for Research and Teaching"
• Lisa C. Orchanian, Thomas P. Smahel, Dianne E. Howie, & Kim J. Vicente

CEL 96-06 "Research on Factors Influencing Human Cognitive Behaviour (III)"
• Dianne E. Howie, Michael E. Janzen, & Kim J. Vicente

CEL 96-07 "Application of Ecological Interface Design to Aviation"
• Nick Dinadis & Kim J. Vicente

CEL 96-08 "Distributed Cognition Demands a Second Metaphor for Cognitive Science"
• Kim J. Vicente

CEL 96-09 "An Experimental Evaluation of Functional Displays in Process Supervision and Control"
• Catherine M. Burns and Kim J. Vicente

CEL 96-10 "The Design and Evaluation of Transparent User Interfaces: From Theory to Practice"
• Beverly L. Harrison

CEL 97-01 "Cognitive Functioning of Control Room Operators: Final Phase"
• Kim J. Vicente, Randall J. Mumaw, & Emilie M. Roth

CEL 97-02 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Two Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 97-03 "ADAPT User's Manual: A Data Analysis Tool for Human Performance Evaluation in Dynamic Systems"
• Xinyao Yu, Farzad S. Khan, Elfreda Lau, Kim J. Vicente, & Michael W. Carter

CEL 97-04 "Research on the Characteristics of Long-Term Adaptation"
• Xinyao Yu, Renée Chow, Greg A. Jamieson, Rasha Khayat, Elfreda Lau, Gerard L. Torenvliet, Kim J. Vicente, & Michael W. Carter

CEL 98-01 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 98-02 "Building an Ecological Foundation for Experimental Psychology: Beyond the Lens Model and Direct Perception"
- Kim J. Vicente

CEL 98-03 "Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-based Work"
- Kim J. Vicente

CEL 98-04 "Ecological Interface Design for Petrochemical Processing Applications"
- Greg. A. Jamieson & Kim J. Vicente

CEL 98-05 "The Effects of Spatial and Temporal Proximity of Means-end Information in Ecological Display Design for an Industrial Simulation"
- Catherine M. Burns

CEL 98-06 "Research on Characteristics of Long-term Adaptation (II)"
- Xinyao Yu, Gerard L. Torenvliet, & Kim J. Vicente