Giving Ergonomics Away?
The application of ergonomics methods by novices

Neville A. Stanton and Mark S. Young*
Department of Design, Brunel University
Runnymede Campus, Egham, Surrey, TW20 0JZ, UK

*now at: Rail Safety and Standards Board,
Evergreen House, Euston Road, London, NW1 2DX

Abstract

A re-occurring theme in applied ergonomics is the idea of "giving the methods away" to those with little formal education in the subject. Little is known, however, about the reliability and validity of these methods when applied to the design process, for novices or experts. It is important to establish just how well the methods will perform in the hands of the analyst. The study reported in this paper presents data on novice intra-analyst and inter-analyst reliability together with criterion-referenced validity across a range of methods. Considerable variation in the reliability and validity of the methods was found. The data were then used in utility analysis, to determine the cost-effectiveness of the methods for an example of car radio-cassette design. The analysis shows that estimates of cost effectiveness may help in the selection of methods.

KEYWORDS: Ergonomics methods, Reliability, Validity, Utility

Ergonomics Methods in Design

Everyone would like well-designed consumer products, from designers, to manufacturers, to consumers. Ease-of-use of a product, the so-called *usability* of a device, is increasingly important as consumers become less accepting of poor design and devices become, potentially, more complex to operate. From discussion with a variety of people in companies involved in the design of consumer products, it became clear that they have a very specific problem of attempting to make decisions on human performance with devices very early in the design process. Often this occurs implicitly by people with no specialist training in ergonomics. In the initial phases of device design, a number of concepts are presented by the design department. From these concepts, a few are chosen as potential products and further investigations into their feasibility are conducted from a variety of viewpoints. From this analysis, one design emerges as the concept crystallises into a definite specification for a device. The earlier that ergonomics can contribute to this process of crystallisation, the more effective the guidance will be in identifying the most appropriate design.

Despite the proliferation of ergonomics methods in research, teaching, and industrial practice, there is little substantive empirical evidence that these methods actually work (Stanton & Young, 1999a). In this respect, there is much that the methods could benefit from the rigor employed in other areas of science. No-one would wish to use a scientific method or instrument without having some idea of how reliable or valid the output is. Why does the ergonomics community seemingly put up with the lack of evidence? The applied psychology community has been engaging in determining the robustness of personnel selection methods for many years. These studies often lead to the expression of inter-analyst reliability and criterion-referenced validity. A review of major texts in the ergonomics domain shows that there is virtually nothing reported by the way of validation studies (e.g., Diaper, 1989a; Kirwan & Ainsworth, 1992; Kirwan, 1994, Corlett & Clarke, 1995; Wilson & Corlett, 1995; Jordan, Thomas, Weerdmeester & McClelland, 1996; Salvendy, 1997; Stanton, 1998; Karwowski, 2001). Ergonomics methods could benefit from a more unambiguous expression of their performance. It is clear that the design community is fairly sceptical about the value of ergonomics, although they can see some merits in the general approach. Therefore, it is important to demonstrate the cost-effectiveness of an ergonomics intervention in design. In this way the reliability and validity of methods are inextricably linked to their utility.

Deleted: 1999

In justifying ergonomics interventions in design, the case could be made more persuasive if it were possible to quantify the cost:benefit ratio. In the domain of personnel selection, researchers have been investigating the utility of methods (Boudreau, 1983). Such research has focused on individual differences in productivity (Schmidt and Hunter, 1983), recruitment decisions (Boudreau and Rynes, 1985), improvements in the utility analysis (Raju, Burke and Normand, 1990; Becker and Huselid, 1992) and ways of presenting utility analysis information to decision makers (Hazer and Highhouse, 1997; Carson, Becker and Henderson, 1998). Ergonomics researchers and practitioners may find their case more convincing if they could express their interventions in terms of overall utility. There have been some attempts to quantify interventions. These normally centre around cost:benefit analyses (e.g. Bias and Meyhew, 1994). This usually works by calculating the cost of applying the method (in terms of person-hours, materials, etc.) and subtracting this from the estimated savings generated by the consequently improved design. The net figure is proposed as the benefit brought about by using the ergonomics methods. They fail, however, to take account of the accuracy of the methods they are using - potential benefits will be reduced if the method is not wholly accurate. It is also recognised that the way in which this information is presented to the decision-maker has an effect upon their willingness to take advantage of it (Hazer and Highhouse, 1997). Carson, Becker and Henderson (1998) argue for greater user-friendliness in presenting utility analysis data. This problem is all too familiar to ergonomics researchers and practitioners. It is simply not sufficient to produce a bottom line figure. Rather, a clear picture of how the data are derived together with an unambiguous interpretation are required.

A survey of the use of ergonomics methods by practitioners in the field reported in this journal, showed that people are unaware of any evidence of published studies of reliability or validity for the methods they were using (Stanton & Young, 1998). The choice of preferred method by the respondents tended to rely upon the few they were used to rather than formal analysis of appropriateness (e.g., reliability or validity) or cost effectiveness (e.g., utility analysis). In the absence of such data to support these analyses (Stanton & Young, 1997), their heuristic approaches to method selection are not surprising. Some researchers have even attempted to formalise the method selection heuristics (e.g., Stanton & Baber, 1996). Research studies of the performance of methods are required if we can expect the selection of methods to become more formal. Stanton & Young (1998) have already reported studies of training and execution times for a range of ergonomics methods. This research needs to be extended further.

With these issues in mind, there are two main aims for this paper. First, we present a validation study of ergonomics methods when applied by non-experts to the evaluation of a typical product. People from other domains will have a tendency to use ergonomics methods if they seem accessible. A recurrent theme in ergonomics research and practice, has been the idea that through training in ergonomics methods we can help non-ergonomists incorporate ergonomic considerations in design (Diaper, 1989b; Wilson, 1995). Training courses in ergonomics methods have been expressly set up with this goal in mind. Training novices in the methods from scratch has the benefit of controlled exposure to the methods, so we can know exactly how much training the participants of the study have had. Also, as Wilson (1995) points out, people from other professions use ergonomics methods and contribute to the overall ergonomics effort. Many of these people might be attracted by the apparent structure of ergonomics methods, some of which have been developed from engineering methods in the first place. Given that the research was partially supported by the Ford Motor Company, we focused on the assessment of car radio-cassette machines. For the second aim, data from this validation study were applied to the analysis of utility in product design scenarios. The combination of these two aims is to provide examples of proof and application of applied psychological research. In order to meet these aims, it was first necessary to train people to use the methods.

Ergonomics Methods

Ten methods were selected for assessment based upon our analysis that these are a representative spread of methods that are currently being used to evaluate human-machine performance and that they were appropriate for the analysis of in-car devices. Methods selected for analysis with a brief explanation of each were as follows:

- Heuristics
- Checklists
- Observation
- Interviews
- Questionnaires
- Link analysis
- Layout analysis
- Systematic Human Error Reduction and Prediction Approach (SHERPA)
- Repertory grids
- Keystroke Level Model (KLM)

Heuristics (Nielsen, 1992)

Heuristics require the analyst to use their judgement, intuition and experience to guide them on product evaluation. This method is wholly subjective and the output is likely to be extremely variable. In favour of the heuristic approach is the ease and speed with which it may be applied. Several techniques incorporate the heuristic approach (e.g., checklists, guidelines, SHERPA) but serve to structure heuristic judgement.

Checklists (Ravden & Johnson, 1989; Woodson, Tillman & Tillman, 1992)
Checklists and guidelines are a useful aide mémoire, to make sure that the full range of ergonomics issues have been considered. However, the approach may suffer from a problem of situational sensitivity, i.e., the discrimination of an appropriate item from a non-appropriate item largely depends upon the expertise of the analyst. Nevertheless, checklists offer a quick and relatively easy method for device evaluation.

Observation (Drury, 1995; Kirwan & Ainsworth, 1992; Baber & Stanton, 1996a)
Observation is perhaps the most obvious way of collecting information about a person's interaction with a device; watching and recording the interaction will undoubtedly inform the analyst of what occurred on the occasion observed. Observation is also a deceptively simple method, one simply watches, participates in, or records the interaction. However, the quality of the observation will largely depend upon the method of recording and analysing the data. There are concerns about the intrusiveness of observation, the amount of effort required in analysing the data and the comprehensiveness of the observational method.

Interviews (Sinclair, 1995; Kirwan & Ainsworth, 1992)
Like observation, the interview has a high degree of ecological validity associated with it: if you want to find out what a person thinks of a device, you simply ask them. Interviewing has many forms, ranging from highly unstructured (free-form discussion) through focused (a situational interview), to highly structured (an oral questionnaire). For the purposes of device evaluation, a focused approach would seem most appropriate. The interview is good at addressing issues beyond direct interaction with devices, such as the adequacy of manuals and other forms of support. The strengths of the interview are the flexibility and thoroughness it offers.

Questionnaires (Brooke, 1996)
There are few examples of standardised questionnaires appropriate for the evaluation of consumer products. However the Software Usability Scale (SUS) may, with some

minor adaptation, be appropriate. SUS comprises 10 items which relate to the usability of the device. Originally conceived as a measure of software usability, it has some evidence of proven success. The distinct advantage of this approach is the ease with which the measure may be applied. It takes less than a minute to complete the questionnaire and no training is required.

Link Analysis (Stammers, Carey & Astley, 1990; Kirwan & Ainsworth, 1992 Drury, 1995)
Link analysis represents the sequence in which device elements are used in a given task or scenario. The sequence provides the links between elements of the device interface. This may be used to determine if the current relationship between device elements is optimal in terms of the task sequence. Time data recorded on duration of attentional gaze may also be recorded in order to determine if display elements are laid out in the most efficient manner. The link data may be used to evaluate a range of alternatives before the most appropriate arrangement is accepted.

Layout Analysis (Easterby, 1984)
Layout analysis builds on link analysis to consider functional groupings of device elements. Within functional groupings, elements are sorted according to the optimum trade-off of three criteria: frequency of use, sequence of use and importance of element. Both techniques (link and layout analysis) lead to suggested improvements for interface layout.

Systematic Human Error Reduction and Prediction Approach (Embrey, 1983; Stanton, 1995; Baber & Stanton, 1996b)
Systematic Human Error Reduction and Prediction Approach (SHERPA) is a semi-structured human error identification technique. It is based upon Hierarchical Task Analysis (HTA) and an error taxonomy. Briefly, each task step in HTA is taken in turn and potential error modes associated with that activity are identified. From this the consequences of those errors are determined. SHERPA appears to offer reasonable predictions of performance but may have some limitations in its comprehensiveness and generalisability.

Repertory Grids (Kelly, 1955; Baber, 1996)
Repertory grids may be used to determine people's perception of a device. In essence, the procedure requires the analyst to determine the elements (the forms of the product) and the constructs (the aspects of the product that are important to its operation). Each

version of the product is then rated against each construct. This approach seems to offer a way of gaining insight into consumer perception of the device, but does not necessarily offer predictive information.

Keystroke Level Model (Card, Moran & Newell, 1983)
The Keystroke Level Model (KLM) is a technique that is used to predict task performance time for error-free operation of a device. The technique works by breaking tasks down into component activities, e.g., mental operations, motor operations and device operations, then determining response times for each of these operations and summing them. The resultant value is the estimated performance time for the whole operation. Whilst there are some obvious limitations to this approach (such as the analysis of cognitive operations) and some ambiguity in determining the number of mental operations to be included in the equation, the approach does appear to have some support.

Study of Reliability and Validity

The objective way to see if the methods work is to assess their reliability and validity. If the methods can be found to be both reliable and valid, they may be used with confidence. The reliability of the methods was assessed in two ways. Intra-analyst reliability was computed by comparing the output generated by each participant at time one with the output at time two. Inter-analyst reliability was computed by looking at the homogeneity of the results of the analysts at time one and at time two. In essence, the purpose of the validity study was to determine the extent to which the predictions were comparable to the actual behaviour of drivers when interacting with the radio-cassette. Criterion-referenced validity was determined by comparing predicted behaviour with actual behaviour in operating the radio-cassette machine.

Method for Reliability Study

Eight male participants and one female participant (who later dropped out of the study) were recruited from the Faculty of Engineering at the University of Southampton. The age range of participants was from 19 to 25 years. Participants were asked to sign a consent form. Engineers were chosen as being representative of the target user population for the outcome of the research.

Design

All participants experienced all methods in the training, practice and application sessions.

Procedure

The procedure contained two main phases, training in methods (in the first week) and application of the methods (in the second week and fourth week) to the evaluation of a device. There was a one week gap between the two application sessions to reduce the effects of common method variance (i.e., participants simply remembering what they did on the first occasion).

Training session in ergonomics methods

In the first week, participants spent up to a maximum of four hours training per method, including time for practice. The training was based upon tutorial notes for training ergonomics methods developed by the authors. The training for each method consisted of an introduction to the main principles, an example of applying the method by case study, and the opportunity to practice applying the method on a simple device. In order to be consistent with other training regimes in ergonomics methods, the participants were split into small groups. In this way they were able to use each other for the interviews, observations, etc. At the end of the practice session each group presented their results back to the whole group and experiences were shared. Timings were recorded for training and practice sessions.

Test sessions applying ergonomics methods

In the second and fourth weeks participants applied each method in turn to the device under analysis. Timings were taken for each method and subjective responses to the methods were recorded on a questionnaire on both occasions. These data are reported by Stanton & Young (1998).

Following the test sessions, participants were thanked for their time and paid for participating in the study. Further details of the training study may be found in Stanton & Young (1998).

Materials

An ergonomics methods training manual (see Stanton & Young, 1999b) was developed to train participants and was accompanied by overhead transparencies during the training session. Training was based upon a SHARP radio-cassette (see Stanton &

Young, 1999b) because this was a similar type of device intended for the application sessions. Participants were allowed to use the training manual during the application sessions. For the purpose of applying the methods to the evaluation of a device, nine radio-cassette machines (Ford 7000 RDS EON - see Stanton & Young, 1999b) were set up in a laboratory.

Method for Validation Study

Participants

Thirty participants (17 males and 13 females), all of whom held a full UK driving licence, were recruited from the University of Southampton. The age range of participants was from 19 to 43 years with a mean age of 25 years. Driving experience ranged from one to 19 years with a mean of 7 years. Annual mileage ranged from 1,000 to 20,000 miles with a mean of 5650 miles. Ethical permission was sought and granted from the Department of Psychology's ethical committee. Participants were asked to sign a consent form and the study complied with BPS ethical standards.

Design

All participants were exposed to all the tasks (listed below) on two trials. The tasks were as follows:

1. Switch radio-cassette on
2. Adjust volume
3. Adjust bass
4. Adjust treble
5. Adjust balance
6. Choose new preset station
7. Choose new station using Seek and store it
8. Use Autostore to store 6 stations
9. Choose a new station using Manual search and store it
10. Use PTY to select a new station
11. Engage News and TA functions
12. Insert cassette
13. Find next track on the other side of cassette
14. Pause cassette to listen to the radio
15. Re-engage the cassette
16. Use AMS to find the next track

17. Engage Dolby NR
18. Eject cassette
19. Switch off

The first trial was considered a learning trial. All the participants had read the manual before commencing. The second trial commenced after a demonstration of the radio functions was given, and the output from this trial was used as a basis for validating the predictions.

Southampton Driving Simulator

The Southampton Driving Simulator was used as the experimental environment, as a car radio-cassette machine was being tested. The simulator comprises an Archimedes RISC computer running simulation software, an Epson colour projection monitor, a projection screen and the front portion of a Ford Orion. The car's controls are fitted with transducers that communicate the driver's actions to the simulator software which alters the viewed image accordingly. The simulation is fully interactive: the driver has full vehicle control and may interact with other vehicles on the road. The data logged include: speed, position on the road, distance from other vehicles, steering wheel and pedal positions, overtakes and collisions. Further details may be found in Stanton, Young & McCaulder (1997).

Procedure

The participants were introduced to the driving simulator and the laboratory. After making themselves comfortable in the car, participants drove the simulator for a few minutes to familiarise themselves with the controls and responses of the computer. The purpose of the experiment was then described to them. Participants were asked to spend as much time as they needed familiarising themselves with the relevant sections of the radio manual. Time spent reading the manual was recorded. The experimental phase began when participants felt comfortable with the workings of the driving simulator and of the radio.

There were two test sessions, each lasting 15 minutes. During these trials, participants were asked to perform the tasks listed above in the order given. Verbal commands were given at regular intervals to facilitate this. Participants were requested to inform the experimenter when they believed they had finished each task, in order that accurate timings, and observed error, could be recorded. Other than the radio tasks, participants

were requested to drive normally and safely, obviously investing their primary attentional resources in driving rather than operating the radio. Between the two test sessions, participants were given a full demonstration of the capabilities of the radio by the experimenter.

After the test sessions were over, participants were asked to give their opinions on usability of the device, in order to validate the predictions made by the engineers. They were therefore given the questionnaire (SUS) to complete, followed by a comprehensive interview which attempted to capture aspects of the methods not already covered by time and error data (e.g., Link and Layout Analysis; Repertory Grids). Following this interview, participants were debriefed, thanked for their time and paid £20 for their participation.

The driving task was performed on a basic track in the Southampton Driving Simulator. The track was based on a figure of eight configuration, modified to provide a range of curves and straight segments (9 straights; 5 right-hand turns;
and 6 left-hand turns). The track distance was 5379m, and all participants completed at least 2 laps in any single run. There were 9 other cars on the track; 6 travelling in the same direction as the user car, and 3 travelling in the opposite direction (a single carriageway set-up). The velocity of these other cars ranged from 30mph to 50mph. Finally, there were 6 roadside objects on the track, including 4 trees and 2 speed limit signs. Participants were instructed to drive as if on a regular journey at their normal speed. Data were recorded on driving speed, road position and headway.

Materials

Participants sat in the Southampton Driving Simulator fitted with a Ford 7000 RDS EON radio-cassette. Driver behaviour was recorded by a miniature camera onto VHS video tape. An interview proforma was used to elicit information from participants. In addition, participants were asked to complete the SUS questionnaire.

Data Reduction

For the purposes of validation, all observed errors and times associated with tasks were noted. Only deviations from an expected task path (defined according to the radio operation manual) were noted as errors. The error data were categorised according to task, and reduced to unique occurrences (i.e., two identical errors associated with the

same task were classified as one error). Frequencies of errors were recorded, however these do not affect the signal detection analysis.

Time data on both trials were recorded; average times for tasks on error-free trials (for KLM), and across all trials (for other analyses), were noted. The post-task interview was designed to glean information pertaining to Heuristics, Checklists, Interviews, Layout Analysis, and Repertory Grids. Thus again all unique pieces of information were noted for purposes of validation by signal detection theory (i.e., repeated occurrences of the same data were not counted, as this artificially inflate the SI statistic). The observational method and KLM both generated time data. It was relatively easy to correlate the time from the participants' performance to the predictions made by the analysts. The other methods required a more sophisticated approach. The signal detection paradigm was adopted to distinguish between appropriate and inappropriate predictions. The predictions from each analysis were classified into one of the four mutually exclusive categories:

- Hits: correctly predict the observed behaviour
- Miss: fail to predict the observed behaviour
- False Alarm: prediction of behaviour not observed
- Correct rejection: correctly reject the behaviour not observed

Then the sensitivity index (SI) was calculated for each participant at time 1 and time 2 as follows:

$$\frac{\left(\frac{\text{Hit}}{\text{Hit} + \text{Miss}} + \left(1 - \frac{\text{False Alarm}}{\text{False Alarm} + \text{Correct Rejection}}\right)\right)}{2}$$

The left hand side of the equation refers to the 'hit rate' (i.e., the rate of correctly identified behaviours) and the right hand side of the equation refers to the 'false alarm' rate (i.e., the rate of incorrectly identified behaviours). Both of these ratios contribute to the overall sensitivity of the method under scrutiny. Ideally the SI should be above 0.5, as this is where the 'hit rate' exceeds the 'false alarm rate'. The closer SI is to 1, the more

accurate the prediction. A discussion of this approach is presented by Stanton & Stevenage (1998).

Analysis

The data were analysed in different ways. First, intra-analyst reliability was determined by using Pearson's correlation coefficient. This measures the degree of consistency of each analyst at time one compared with the same analyst at time two. Second, inter-analyst reliability was computed using the Kurtosis statistic. This looks at the degree to which ratings are spread with each group of analysts at time one and time two separately. Finally, validity was analysed by assessing the value of SI at time one and time two. This value is the combination of the *hit rate* and *false alarm* rate. The distinction is an important one, because it is as important to predict true positives as it is to reject false positives.

Observation, Questionnaires and KLM were treated differently for some of the analyses. Two correlation coefficients were computed for the intra-analyst reliability of the observations, one for error data and one for time data. Correlation coefficients were computed for the predictive validity of the Questionnaires and KLM instead of SI as there was a directly comparable quantitative measure (i.e., the questionnaire was completed by participants and time data were available for the study of KLM).

Results and Discussion of Reliability and Validity Study

Statistical differences were found in two of the three driving measures. In the second trial drivers positioned the vehicle slightly closer to the centre line than in the first trial ($t = -2.2$, $p<0.05$). Drivers also drove slightly faster in the second trial compared to the first ($t = -3.55$, $p<0.01$). There were no differences in headway in the two trials ($t = 0.86$, p=NS). These results could be due either to the radio-cassette tasks being less intrusive on the driving task in the second trial or to a slight improvement in driver performance on the second trial. Whichever of these effects accounts for the difference in performance has no real bearing on the main purpose of the study, which was to assess the reliability and validity of the methods.

The data analysis of reliability and validity are presented in table one. The reliability and validity data are presented together because the two concepts are inter-related. Whilst a method might be reliable (i.e., it might be stable across time and/or stable across analysts) it might not be valid (i.e., it might not predict behaviour). However if a

method is not reliable it cannot be valid. Therefore the relationship between reliability and validity can be said to be unidirectional.

Table 1. Summary of reliability and validity statistics for Ergonomics methods based on a study of novices

Methods	Intra-analyst Reliability	Inter-analyst Reliability	Concurrent Validity
Heuristics	r=0.471 (p=ns) Z=-1.13 (p=ns)	K=0.777 (T1) K=-1.27 (T2)	SI=0.464 (T1) SI=0.476 (T2)
Checklists	r=0.307 (p=ns) Z=-1.13 (p=ns)	K=4.40 (T1) K=2.52 (T2)	SI=0.602 (T1) SI=0.587 (T2)
Observation: errors: time:	r=0.890 (p<0.005) Z=-1.35 (p=ns)	K=-0.177 (T1) K=0.180 (T2)	SI=0.466 (T1) SI=0.474 (T2) r=0.729 (p<0.001) t=3.68 (p<0.001)
Interviews	r=0.449 (p=ns) Z=-1.52 (p=ns)	K=-1.66 (T1) K=0.362 (T2)	SI=0.488 (T1) SI=0.466 (T2)
Questionnaires	r=0.578 (p=ns) Z=-1.35 (p=ns)	K=-0.908 (T1) K=0.812 (T2)	r=0.563 (p=ns) (T1) Z=-2.80 (p<0.01) (T1) r=0.615 (p=ns) (T2) Z=-2.80 (p<0.01) (T2)
Link Analysis	r=0.830 (p<0.05) Binomial (p=ns)	K=-1.42 (T1) K=0.075 (T2)	SI=0.758 (T1) SI=0.764 (T2)
Layout Analysis	r=-0.121 (p=ns) Z=-1.07 (p=ns)	K=-0.152 (T1) K=0.841 (T2)	SI=0.041 (T1) SI=0.070 (T2)
SHERPA	r=0.392 (p=ns) Binomial (p=ns)	K=-1.37 (T1) K=1.68 (T2)	SI=0.628 (T1) SI=0.614 (T2)
Repertory Grids	r=0.562 (p=ns) Z=-0.00 (p=ns)	K=-0.0195 (T1) K=-0.710 (T2)	SI=0.519 (T1) SI=0.533 (T2)
KLM	r=0.916 (p<0.001) Z=-0.355 (p=ns)	K=0.522 (T1) K=2.91 (T2)	r=0.890 (p<0.001) (T1) Z=-3.43 (p<0.001) (T1) r=0.769 (p<0.001) (T2) Z=-3.58 (p<0.001) (T2)

In addressing intra-analyst reliability, three of the methods achieved acceptable levels, denoted by the statistically significant correlations. These methods were:

- Observation
- Link Analysis
- Keystroke Level Model

This means that the analysts' predictions were stable across time.

Two methods achieved acceptable levels of inter-analyst reliability, as evidenced by the Kurtosis statistic (which is an indicator of how closely grouped the analysts predictions were to each other), where a value of greater than zero means that the data are steeper (therefore more tightly grouped) than the normal distribution curve and a value of less than zero means that the data are flatter (therefore more distributed) than the normal distribution curve. Ideally, values should be greater than zero to indicate greater agreement between analysts. The more positive the value, the greater the degree of agreement. Generally speaking, the values improved between time one and time two, suggesting that the analysts were learning how to apply the techniques. Methods that performed at an acceptable level (at time two) are as follows:

- Checklists
- Keystroke Level Model

This means that the methods listed above showed an acceptable level of agreement between analysts. Finally, criterion-referenced validity was computed from SI, with the exception of the observation, questionnaires and KLM (where Pearson's correlation coefficient was used). A value of greater than 0.5 for SI was the criteria for acceptance of the method (or a statistically significant correlation in the case of observation, questionnaires and KLM). Methods that performed at an acceptable level were as follows:

- Checklists
- Observation (time data only)
- Link Analysis
- Systematic Human Error Reduction and Prediction Approach
- Repertory Grids

- Keystroke Level Model

This means that the methods listed above seemed to capture some aspects of the performance of the participants engaged in the study of the radio-cassette. However, as pointed out earlier, validation data cannot be interpreted independently of reliability data. Therefore only one of the methods performed at an acceptable level for all three criteria:

- Keystroke Level Model

Relaxing these criteria a little would allow us to consider five more methods that performed at an acceptable level with respect to criterion-referenced validity (with the proviso that the evidence suggests that the methods may not stable either over time or between analysts), these are:

- Link Analysis
- Checklists
- Systematic Human Error Reduction and Prediction Approach
- Observation
- Questionnaires

Given that methods cannot be valid unless they are proven to be reliable also, and there is little point in using methods that are reliable unless they are proven to be valid, we recommend that all of the other methods are treated with caution until further studies have established their reliability and validity.

Utility Analysis

To determine the relative benefits of these methods when applied in the field, a utility analysis equation has been derived to assign an approximate financial value to each method. The equation is based on the accuracy of each method, the cost of retooling or redesigning (i.e., "TGW", or "things gone wrong"), and the cost in person hours of using the technique.

Accuracy

From the validation study, data have been acquired on inter-analyst reliability (how consistently different people use the method), intra-analyst reliability (how consistently the same person uses the method on different occasions), and validity (how well the method predicts what it is supposed to). The coefficients of each of these variables have been transformed where necessary so that they lie between 0 and 1. This enables a simple multiplication to provide information on the overall accuracy of the method (which will also be between 0 and 1). That is, given a random analyst at any point in time, how well can s/he be expected to perform with a given method? Accuracy can therefore be summarised thus:

accuracy = inter-analyst reliability * intra-analyst reliability * validity

TGW Costs

The study referred to above was based on an in-car stereo system, so the retooling figures we refer to here are also based on a car radio. Retooling costs were supplied to us by our colleagues at Ford Motor Company as the most significant aspect of the redesign process. Of course, the analyst can substitute these figures with their own if they are interested in a different product. Retooling costs for a car radio can be between £3000 (for minor changes to the product line) and £150000 (for complete retooling). Assuming that the accuracy of a method represents how much of these things gone wrong could be saved, multiplying the accuracy by the retooling costs will reveal how much money each method will yield:

savings = accuracy * retooling costs

Costs of using the method

Of course, this isn't the final figure, because there are costs involved in using the technique. If we assume an analyst is worth £50 per hour, each hour spent using the method will mean £50 less savings. Therefore our total utility for each method is:

utility = (accuracy * retooling costs) - method costs

Substituting the above equations into this one provides us with the final utility equation:

Utility = $(r_1 * r_2 * v * C_t) - C_m$
where : r_1 = inter-analyst reliability
r_2 = intra-analyst reliability
v = validity
C_t = retooling costs
C_m = costs of using the method

Using the equation
There are four aspects of ergonomics which the methods attempt to predict: errors, performance times, usability, and task sequence. Most of the methods fit best into just one of these categories; the exception being Observation, which can be used to predict both errors and performance times. The relationships between method and output are summarised in table two.

Table 2. Output from the 12 methods

Output			
Errors	**Times**	**Usability**	**Task sequence**
SHERPA	KLM	Checklists	Link Analysis
Observation	Observation	Questionnaires	Layout Analysis
		Repertory Grids	
		Interviews	
		Heuristics	

Given these four areas, it could be assumed that they each account for an equal proportion of retooling costs. This is probably an oversimplification, but it is

only meant to be used for an heuristic analysis. Allowing for a similar sized proportion for residual error, that would mean each area accounts for 20% of the retooling costs for a device. So, the first step in using the equation is to divide the retooling costs by 5. Retooling costs will be specific to each situation, and this variable needs to be adjusted as appropriate. Similarly, analyst costs (Cm) can also be adjusted for more or less expensive consultants. The rest of the variables for the equation are summarised in table three.

Table 3. Reliability, validity and costs associated with each method based on novices.

Utility = (r_1 * r_2 * v * C_t) - C_m				
Method	**r_1**	**r_2**	**v**	**C_m (£)**
KLM	0.754	0.916	0.769	112.5
Link Analysis	0.286	0.830	0.764	104.2
Checklists	0.690	0.307	0.587	83.3
SHERPA	0.551	0.392	0.614	241.7
Observation	0.304	0.890	errors: 0.474 times: 0.729	125
Questionnaires	0.408	0.578	0.615	37.5
Repertory Grids	0.157	0.562	0.533	112.5
Layout Analysis	0.413	0.121	0.070	70.8
Interviews	0.334	0.449	0.466	283.4
Heuristics	0.0644	0.471	0.476	62.5

N.B. The cost of the methods (C_m) is based on time taken to analyse a car radio, and includes training and practice time. Note also that Observation has two validity statistics associated with it, depending on whether errors or performance times are of primary concern.

Worked example

Here are two examples of using the utility equation to demonstrate the payoff of using a particular method on a car radio. The first demonstrates using the best method, KLM.

Step 1: Calculate retooling costs

A conservative estimate for retooling costs involved with a car radio could be set at £5000. KLM covers one area of ergonomics - performance times - thus can at best be expected to account for 20% of this, or £1000.

Step 2: Insert variables into the equation

Utility = $(r_1 * r_2 * v * C_t) - C_m$

U_{KLM} = (0.754 * 0.916 * 0.769 * 1000) - 112.5
= 531.1 - 112.5
= **£418.6**

So, using KLM before commissioning this product could save about £420 on minor retooling costs.

Compare these figures with those obtained when Heuristics are used:

Utility = $(r_1 * r_2 * v * C_t) - C_m$

$U_{Heuristics}$ = (0.0644 * 0.471 * 0.476 * 1000) - 62.5
= 14.4 - 62.5
= **-£48.1**

Here, the costs of using the method outweigh the benefits. Of course, if the potential retooling costs were higher, the savings would be too, and this picture may well be different.

Using more than one method

In some cases, being restricted to one technique would be a disadvantage. How can utility be calculated for 2 or more techniques?

a) The methods assess different aspects of ergonomics

If the chosen methods lie in separate categories of those outlined above, then simply calculate the utility for each method separately and sum the amounts at the end. For a simple example, take the two methods already calculated. KLM assesses performance

times, and Heuristics is concerned with design. Of the £5000 total retooling costs, £1000 of this could be due to performance times, and a further £1000 due to design. So the total maximum potential saving is £2000.

Combined utility = $U_{KLM} + U_{Heuristics}$ = 418.6 + (-48.1) = **£370.5**

b) The methods assess the same aspect of ergonomics

This situation is slightly more complex. Because 20% of the retooling costs are allocated to each area, this proportion has to be shared somehow. Assume that the methods will be executed in order of accuracy, best first. Calculate the **savings** (not overall utility) for the first method. Then perform the utility analysis for the second method on the **remainder**. Sum the respective utilities at the end of this process and you have the overall utility for using the methods in combination. Staying with the KLM example, let's say it is to be used with Observation to predict performance times. The savings generated by KLM (before subtracting the costs of using the method) are £531.1, leaving £468.9 out of the original £1000. Now use the utility equation for Observation on this £468.9 (be aware to insert the correct validity statistic for Observation predicting performance times):

For Method 1 (KLM):

Savings = $r_1 * r_2 * v * C_t$	**Utility = $(r_1 * r_2 * v * C_t) - C_m$**
S_{KLM} = 0.754 * 0.916 * 0.769 * 1000	U_{KLM} = (0.754 * 0.916 * 0.769 * 1000) - 112.5
= 531.1	= 531.1 - 112.5
Remainder = 1000 - 531.1 = **468.9**	= **£418.6**

For Method 2 (Observation):

Utility = $(r_1 * r_2 * v *$ Remainder$) - C_m$

$U_{Obs.}$ = (0.304 * 0.890 * 0.729 * 468.9) - 125

= **-£32.5**

overall utility of using both KLM and Observation = 418.6 + (-32.5) = £386.1

Summary

The utility equation described here is intended to provide an approximate insight into how much money each method is potentially worth to designers and engineers. It is purely a cost-benefit tool, and not intended to be so accurate as to be used in accounting. The reader should also be aware that the method costs (C_m) are based on analysing a car radio, so may change with other devices. Retooling costs will also differ, so it is up to the analyst to substitute these accordingly. It is recommended that a conservative attitude is adopted for this. However, these issues aside, the utility analysis can provide a tangible forecast about the usefulness of each method, albeit approximate. It may also aid choosing between methods, for the relative advantages of one method over another may be more clearly defined in monetary terms.

Conclusions

In conclusion, this paper has sought to present data on the inter-analyst and intra-analyst reliability and criterion-referenced validity of ergonomics methods applied to the design evaluation process. This is, as far as we know, the first time that a study has sought to quantify these methods in this way. The data show a far from wholesome picture and would suggest caution, particularly when ergonomics methods are in the hands of a novice analyst. The study of reliability and validity favours KLM, link analysis, checklists, and SHERPA. It is not by chance that the top performing methods in terms of reliability and validity concentrate on very narrow aspects of performance (i.e., mainly the observable actions). Generally speaking, the broader the scope of the analysis the more difficult it is to get favourable reliability and validity statistics. This does not negate the analysis, however. We are arguing that analysts should be aware of the potential power of the method *before* they use it, rather than proposing that they should not use it. It is an important goal of future research to further establish the reliability and validity of ergonomics methods in different contexts. A study of expert users would also be useful as these data might be very different, as would the study of devices with greater and less complexity. The factors of analyst's expertise and device complexity are likely to interact. Stanton & Stevenage (1998) found that novices performed significantly better than those in this study when analysing a much simpler device. By way of contrast, Baber & Stanton (1996b) showed that expert analysts performed significantly better than those in this study when analysing a more complex device. The research debate is likely to continue for some time. It has taken researchers in the field of personnel selection some forty years to reach a general consensus of opinion about the performance of their methods. On this basis, it could be some time before a similar status is achieved for ergonomics methods. The study also shows how

the data may be utilised as part of a cost:benefit trade-off when comparing the intervention strategies that could be employed. Relative, rather than absolute values of utility are likely to be most applicable. The advantage of the system is in its simplicity. This makes it fairly easy to audit the utility equation and is likely to be a factor in gaining its acceptance by designers when justifying an evaluative study.

This research represents an initial start towards the data required by practitioners in the justification for the use of ergonomics methods, and may even assist in the selection of methods for a particular intervention. Caution is urged in using the values in absolute terms for several reasons. As Annett (2002) points out, the data are based on novice users of the methods after a training regime lasting only one week. He argues that the underlying theory supporting the method together with the skill and expertise of the analyst should also be taken into account. It seems reasonable to assume that the greater the expertise of the analyst, the more sensitive their application of the method. It remains an important goal of future research to quantify this improvement in terms of reliability and validity, rather than just accepting that experts will be better than novices.

Two further criticisms have been levelled at this research. These concerns centre on the use of the reliability and validity statistics in the utility analysis equation. The first criticism is that we do not need to factor reliability into the utility analysis equation as "*validity will already be attenuated by unreliability.*" Our reason for including it was that we were trying to stay close the style of utility analysis used for the evaluation of personnel selection methods. In that field the utility analysis formula considers the variation in performance of the potential personnel (called SDy) as well as the validity and cost of the methods in question. We have simply substituted reliability in place of SDy, as it represents the variability in the performance of people using the ergonomics method(s). Whilst we accept that these are not the same things we thought that it captured the spirit of the analysis. We are not, however, going to object to people using the modified version of the formula with reliability statistics removed as this presents a more optimistic value! The second criticism is that we should use a correlation coefficient for the validity statistic rather than SI in the utility analysis equation (Blinhorn, S., 1999, personal communication) To counter this arguement we have calculated Phi correlation coefficients in place of SI and present them in table four together with the Pearson product moment correlation coefficients. In addition, Phi

enables PhiMax to be calculated. PhiMax is the maximum values that theoretically could be achieved with the data.

Table 4. Concurrent validity expressed as correlation coefficients for novice application of ergonomics methods.

Method	**Pearson**	**Phi**	**PhiMax**
Heuristics	-	0.087	0.464
Checklists	-	0.206	0.659
Observation: errors	-	-0.141	0.343
Observation: time	0.729	-	-
Interviews	-	-0.112	0.422
Questionnaires	0.615	-	-
Link Analysis	-	0.356	0.572
SHERPA	-	0.238	0.922
Repertory Grids	-	0.078	0.681
KLM	0.769	-	-

Thus the utility analysis could be simplified thus: $\mathbf{U = (v * C_t) - C_m}$

Some readers might argue that ergonomics methods should only be used by ergonomics experts, but we were trying to push the boundaries of the method and extend its utility in a practical environment. If non-experts pick up the Stanton & Young book (for example), they might be inclined to use the methods regardless of their expertise. We accept that the reader may not approve of this on both counts: first the methods should only be used by experienced ergonomists and second the methods should only be used when there is access to the end-user population. Returning to the research question posed at the beginning of this paper, it seems that some of the simpler forms of analysis by ergonomics methods can indeed be 'given away' to relative novices with appropriate training and supervision. Such methods would include KLM, questionnaires and observation (time). For other methods, greater caution is recommended. At the end of the day, any method is no substitute for ergonomics knowledge and expertise.

Acknowledgements

The research reported in this paper was supported by the Engineering and Physical Sciences Research Council (UK) under the EPSRC/DTI-LINK Transport Infrastructure and Operations Programme. The authors are grateful to the Ford Motor Company for supplying the radio-cassette machines and the Orion car. The authors are also grateful to the reviewers for helping to improve this paper.

References

Annett, J. 2002 A note on the validity and reliability of ergonomics methods. *Theoretical Issues in Ergonomics Science*, 3 (2) 228-232.

Baber, C. 1996 'Repertory grid theory and its application to product evaluation' in P. W. Jordan; B. Thomas; B. A. Weerdmeester & I. L. McClelland (eds) *Usability Evaluation in Industry* Taylor and Francis, London, pp 157-165

Baber, C. and Stanton, N. A. 1996a 'Observation as a usability method' in P. W. Jordan; B. Thomas; B. A. Weerdmeester & I. L. McClelland (eds) *Usability Evaluation in Industry* Taylor and Francis, London, pp 85-94

Baber, C. and Stanton, N. A. 1996b 'Human error identification techniques applied to public technology: predictions compared with observed use' *Applied Ergonomics* 27, 119-131

Becker, B. E. & Huselid, M. A. 1992 Direct estimates of SDy and the implications for utility analysis. *Journal of Applied Psychology*, 77, 227-233.

Bias, R. G and Meyhew, D. J. 1994 Cost-Justifying Usability. Boston: Academic Press.

Boudreau, J. W. 1983 Effects of employee flows on utility analysis of human resource productivity improvement programs. *Journal of Applied Psychology*, 68, 396-406.

Boudreau, J. W. & Rynes, S L. 1985 Role of recruitment in staffing utility analysis. *Journal of Applied Psychology*, 70, 354-366.

Brooke, J. 1996 'SUS: a 'quick and dirty' usability scale' in P. W. Jordan; B. Thomas; B. A. Weerdmeester & I. L. McClelland (eds) *Usability Evaluation in Industry* Taylor and Francis, London, pp 189-194

Card, S. K., Moran, T. P. and Newell, A. 1983 *The Psychology of Human-Computer Interaction*. Erlbaum, Hillsdale NJ.

Carson, K. P., Becker, J. S. & Henderson, J. A. 1998 Is utility really futile? A failure to replicate and an extension. *Journal of Applied Psychology*, 83, 84-96.

Corlett, E. N. & Clarke, T. S. 1995 *The Ergonomics of Workspaces and Machines*. 2nd Edition. Taylor and Francis, London.

Diaper, D. 1989a *Task Analysis in Human Computer Interaction*. Ellis Horwood, Chichester.

Diaper, D. 1989b Giving HCI away. In: A. Sutcliffe and L. Macaulay (eds) People and Computers V. British Computer Society, London. pp. 99-117.

Drury, C. G. 1995 'Methods for direct observation of performance' in J. Wilson and N. Corlett (eds) *Evaluation of Human Work*. 2nd Edition. Taylor & Francis, London, pp 45-68.

Easterby, R. 1984 Tasks, processes and display design in R. Easterby and H. Zwaga (eds) *Information Design* Taylor and Francis, London. pp.???

Embrey, D. 1983 'Quantitative and qualitative prediction of human error in safety assessments' in the Institution of Chemical Engineers symposium Series 130. *I.Chem.E., London*, 329-350.

Hazer, J. T. & Highhouse, S. 1997 Factors influencing manager's reactions to utility analysis: effects of SDy method, information frame, and focal attention. *Journal of Applied Psychology*, 82, 104-112.

Jordan, P. W.; Thomas, B.; Weerdmeester, B. A. & McClelland, I. L. 1996 *Usability Evaluation in Industry*. Taylor and Francis, London.

Karwowski, W. 2001 *International Encycolpedia of Ergonomics and Human Factors*. Taylor & Francis, London.

Kelly, G. A. 1955 *The Psychology of Personal Contructs*. Norton, New York.

Kirwan, B. and Ainsworth, L. 1992 *A Guide to Task Analysis*. Taylor & Francis, London.

Kirwan, B. 1994 *A Guide to Practical Human Reliability Assessment*. Taylor and Francis, London.

Nielsen, J. 1992 'Finding usability problems through heuristic evaluation' in *Proceedings of the ACM Conference on Human Factors in Computing Systems* ACM Press, Monterey CA, pp 373-380.

Raju, N. S., Burke, M. J. & Normand, J. (1990) A new approach for utility analysis. *Journal of Applied Psychology*, 75, 3-12.

Ravden, S. J and Johnson, G. I. 1989 *Evaluating Usability of Human-Computer interfaces: a practical method*. Ellis Horwood, Chichester.

Salvendy, G. 1997 *Handbook of Human Factors and Ergonomics*. Wiley, New York.

Schmidt, F. L. & Hunter, J. E. 1983 Individual differences in productivity: an empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology*, 68, 407-414.

Sinclair, M. 1995 'Subjective assessment' in J. Wilson and N. Corlett (Eds) *Evaluation of Human Work*. 2nd Edition. Taylor & Francis, London, pp 69-100.

Stammers, R. B. Carey, M. and Astley, J. A. 1990 'Task analysis' in J. Wilson and N. Corlett (eds) *Evaluation of Human Work*. Taylor & Francis, London, pp 134-160.

Stanton, N. A. 1995 Analysing worker activity: a new approach to risk assessment *Health and Safety Bulletin*, 240 pp 9-11.

Stanton, N. A. 1998 *Human Factors in Consumer Products.* Taylor and Francis, London.

Stanton, N. A. and Baber, C. 1996 'Factors affecting the selection of methods and techniques prior to conducting a usability evaluation' in P. W. Jordan; B. Thomas; B. A. Weerdmeester & I. L. McClelland (eds) *Usability Evaluation in Industry*. Taylor and Francis, London pp 39-48.

Stanton, N. A. & Stevenage, S. (1998) Learning to predict human error: issues of acceptability, reliability and validity. *Ergonomics*, 41 (11), 1737-1756.

Stanton, N. A. & Young, M. S. 1997 Validation: the best kept secret in ergonomics. In: D. Harris (Ed) *Engineering Psychology and Cognitive Ergonomics volume 2: Job Design and Product Design*. Ashgate, Aldershot pp.301-307.

Stanton, N. A. & Young, M. S. 1998 Is utility in the mind of the beholder? A review of ergonomics methods. *Applied Ergonomics*, 29 (1) 41-54

Stanton, N. A. & Young, M. S. 1999a What price ergonomics? *Nature* 399, 197-198.

Stanton, N. A. & Young, M. S. 1999b *A Guide to Methodology in Ergonomics: Designing for Human Use.* Taylor and Francis, London.

Stanton, N. A., Young, M. S. & McCaulder, B. 1997 _rive-by-wire: the case of mental workload and the ability of the driver to reclaim control. *Safety Science* 27 (2-3) 149-159

Wilson, J. 1995 A framework and context for ergonomics methodology. In: Wilson, J. and Corlett, N. (eds) *Evaluation of Human Work.* 2nd Edition. Taylor and Francis, London. pp. 1-39.

Wilson, J. and Corlett, N. 1995 *Evaluation of Human Work.* 2nd Edition. Taylor and Francis, London.

Woodson, W. E., Tillman, B. and Tillman, P. 1992 *Human Factors Design Handbook* 2nd edition, McGraw-Hill, New York.