Validating ergonomics methods

Neville A Stanton

Human Factors Engineering, Transportation Research Group, School of Civil, Maritime and Environmental Engineering, Faculty of Engineering and Physical Sciences, Boldrewood Innovation Campus, Burgess Road, University of Southampton, SO16 7QF, UK

ABSTRACT

This paper revisits the challenge laid down over 15 years ago, that Ergonomics needs to report on the reliability and validity of its methods if it is to maintain its standing amongst the Engineering community. Unfortunately, a review of books reporting on Ergonomics methods since that time shows very little change. The theoretical constructs of reliability and validity are presented together with the rationale for conducting studies of training in Ergonomics methods. Revisiting the original study shows the way in which such validation work can be undertaken and data reported. It is hoped that this article provides the stimulus for more studies of this nature.

KEYWORDS

Methods, Reliability, Validity

Introduction

In 1999, we threw the gauntlet down to the discipline of Ergonomics, to prove that their methods actually do all that it is claimed of them (Stanton and Young, 1999a, b). Twenty years later, there appears to have been relatively little take-up of that challenge. An analysis of the state-of-the-science shows very little change on reporting of studies (Stanton et al, 2014). In the analysis of Ergonomics methods reported in 15 books published since the original challenge in 1999, there is only a smattering of reports on studies of training and validation of the methods. Few of the texts contain any description of the relative merits of one method over another. Again, we have attempted to redress this position by providing data on both impacts of training people to use the methods as well as data on the reliability and validity of those methods (Stanton and Young, 2003; Stanton et al, 2005a,b; Stanton et al, 2013). It has been argued that all Ergonomics methods need to prove that they can work in the intended domain of applications (Stanton, 2014). For Ergonomics as an Engineering discipline, this does not seem to be an unreasonable request, as it is necessary for both our academic and professional credibility. All methods should have to demonstrate they have met the criteria for both reliability and validity as proposed by Stanton and Young (1999a,b).

Reliability

Reliability is a measure of stability of the method over time and stability of the method across analysts. Ideally, it should be possible to demonstrate that the application of an Ergonomics method will result in the same results if it is used by different people, or on different occasions by the same people (provided that the system being analysed hasn’t changed). A method is generally considered to have minimally acceptable reliability if the method’s expert creator could achieve repeatable results on different occasions. At the other extreme would be a method that delivered the same results when used by anyone, with even a little training. Between these extremes would be most of the methods used by Ergonomists. Whether any one of these would be considered to have an acceptable degree of reliability would depend on a variety of factors, including: the expertise of those using it; various constraints such as time and resource availability; the type of project; and the problem for which the method was being used. Indeed, when a method is being used creatively then high reliability, either for an individual analyst or across different analysts, may be undesirable, as it could restrict the range of alternatives considered. By way of contrast, in large, safety-critical, projects with a number of analysts, a much higher degree of reliability is necessary, as the results achieved by the different analysts will no doubt need integrating at some stage during the project.

Validity

If reliability is a challenging concept, then validity is even more so. Stanton and Young (1999b) proposed four types of validity for Ergonomics methods: construct, content, concurrent, and predictive (see Figure 1). Construct validity, for example, concerns the underlying theoretical basis of a method. Content validity, according to Stanton and Young, is concerned with the credibility that a method is likely to gain among its users. They suggest that, ideally, a method should use appropriate terminology and language and seem up to the job of analysis if it is to be taken seriously. Obviously, such validity requires agreement among those using the methods. Finally, Stanton and Young (1999a, b) argued that concurrent and predictive validity concerns address the extent to which an analysed performance is representative of the performance that might have been analysed. The difference between concurrent and predictive validity is a matter of time: concurrent validity describes current performance sampled whereas predictive validity concerns the performance of the future. What is important is that the Ergonomics methods posses a level of concurrent or predictive validity suitable for their application. There continues to be debate over the role of validation in Ergonomics (Annett, 2002; Stanton, 2002; Stanton, 2014; Stanton & Young, 1999b), and the issues are by no means resolved. The goal of the discipline for methods should be to meet both reliability and validity criteria. Although laboratory and other research work may be a desirable minimum, it is the perceptions of ultimate users in the design and engineering industries that will be most important.

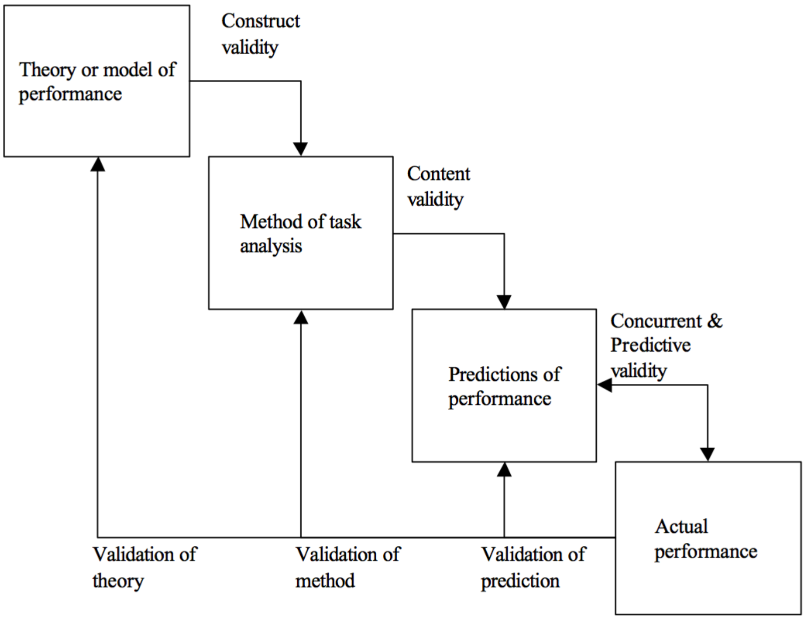

Figure 1. Types of validation of Ergonomics Methods

Examples of studies of validation include Ergonomics methods such as the Systematic Human Error Reduction and Prediction Approach (SHERPA: Stanton and Stevenage, 1998), Task Analysis for Error Identification (TAFEI: Stanton and Baber, 2005), Human Error Template (Stanton et al, 2009) and systems analysis methods such as Cognitive Work Analysis (Cornelissen et al, 2015) as well as a range of methods more generally (Stanton and Young, 2003). In particular, we have pioneered the use of the Signal Detection Paradigm as a means of establishing empirical validity of methods (Stanton and Young, 1999a,b). More recently, we have been arguing for a standard reporting of Ergonomics methods in their development and validation (Stanton, 2014).

Training

Very little has been written on training people to use ergonomic methods, as noted by Stanton and Stevenage (1998). In order to evaluate the ease with which people are able to acquire ergonomic methods, a study into the training and application of each method by novice analysts was reported by Stanton and Young (1999a, b, 2003). In the first week, participants spent up to a maximum of 4 hours training per method, including time for practice. The training was based upon tutorial notes focused on the training of ergonomics methods. 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 ergonomic 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 practise session, each group presented their results back to the whole group and experiences were shared. Timings were recorded for training and practice sessions. In the second and fourth weeks, participants applied each method in turn to the device under analysis. Timings were taken for each method.

These data seem to reinforce the reason for the popularity of questionnaires, interviews, observations, checklists and heuristics noted in the survey (Stanton and Young, 1998a) as they take relatively little time to learn when compared with HTA, SHERPA and TAFEI. Perhaps it is surprising to see that link and layout analysis are not more popular, given that they are also relatively quick to train people in. Similarly, repertory grids and the Keystroke Level Model (KLM) seem to be no more time-consuming to train people in than the focused interview. However, these techniques are rather more specialised in their output, like link and layout analysis.

There were significant differences in the time taken to analyse a device using different methods. The popularity of questionnaires, observations and checklists is reinforced by them being relatively quick and flexible methods. The speed with which observations are conducted is perhaps counterintuitive, but the time taken in execution of the method is for a single participant only. And note that heuristics and interviews appear to take as long as link analysis, repertory grids and Critical Path Analysis (CPA), whereas layout analysis appears quicker. HTA and SHERPA take approximately the same time as each other, but they are much more time-intensive than other methods. Besides that, SHERPA requires the output of HTA, so it would require the time to conduct HTA plus the time to conduct SHERPA if it were to be used in a situation where no HTA had been developed. Over half the participants failed to complete TAFEI within the time available, suggesting that it was the most time-consuming of the methods under test.

Utility

Not only do the methods have to be acceptable to the users, but they also have to work. The objective way to see whether the methods work is to assess their reliability and validity. If the methods can be demonstrated as reliable and valid, they may be used with confidence. The reliability of the methods was assessed in two ways. Intra-rater reliability was computed by comparing the output generated by each participant at trial 1 with the output at trial 2. Correlation coefficients were computed to assess the stability of the measures. Inter-rater reliability was computed by looking at the homogeneity of the results of the analysts at trial 1 and at trial 2. 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 in-vehicle device. The data were analysed in different ways. First, intra-rater reliability was determined by using Pearson's correlation coefficient. This measures the degree of consistency of each rater at trial 1 compared with the same rater at trial 2. Second, inter-rater reliability was computed using the Kurtosis statistic. This looks at the degree of spread for the ratings within each group of raters at trial 1 and then at trial 2. Finally, validity was analysed by assessing the value of d' at trial 1 and at trial 2. This value is the combination of the hit rate and false alarm rate. The distinction is a significant one, because it is as important to predict true positives as it is to reject false positives (within the Signal Detection Theory paradigm).

Reliability and validity was calculated for each method (see table 1) in the study reported by Stanton and Young (1999a,b, 2003). A method might be reliable (i.e. it might be stable across time and/or stable across analysts) but 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 described as unidirectional.

Table 1. Reliability and validity statistics for each method (ranked in order of overall performance – where better performance is closer to 1)

Method	Inter-analyst reliability	Intra-analyst reliability	Validity
KLM	0.754	0.916	0.769
Link analysis	0.286	0.830	0.764
Checklists	0.690	0.307	0.587
SHERPA	0.551	0.392	0.614
Observation:			
Errors	0.304	0.890	0.474
Times	0.209	0.623	0.729
Questionnaires	0.408	0.578	0.615
HTA	0.206	0.226	0.591
Repertory grids	0.157	0.562	0.533
Layout analysis	0.413	0.121	0.070
Interviews	0.334	0.449	0.466
Heuristics	0.0644	0.471	0.476
TAFEI	-	-	0.506

Note: Kurtosis statistics have been transformed to give values between 0 and 1.

The methods in the study by Stanton and Young (1999a,b) did differ considerably in terms of reliability and validity. They suggested that newly trained novices perform better with some methods than others; KLM performed best in the test. The next grouping was for link analysis, checklists and SHERPA. The third grouping was the rest of the methods. Stanton and Young (1999b) urged caution for this latter set of methods, particularly when they are used by novice analysts. Later studies have shown that experts generally yield better reliability and validity data than novices (Stanton and Baber, 2002, 2005).

Conclusions

In conclusion, despite accepted ways of testing being available, there is clearly little reported evidence in the literature on reliability or validity of ergonomics methods. This was confirmed by the survey undertaken by Stanton and Young (1998a) and in a later update (Stanton, 2014). The training study shows that some methods are easier to acquire than others, which is probably related to their relative popularity. The study of reliability and validity also shows that some methods perform better than others, for a novice population at least. It is not by chance that the top-performing methods in terms of reliability and validity concentrate on very narrow aspects of performance (Stanton and Young, 1999a,b; Stanton et al, 2014). Generally speaking, the broader the scope of the analysis, the more difficult it is to get favourable reliability and validity statistics. But this does not negate the usefulness of the analysis. What is being argued is 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. Ergonomists and designers would be well served by exploring the benefits of other methods rather than always relying upon two or three favourite approaches (Stanton and Young, 1998a). Despite little research undertaken over the past twenty years, establishing the reliability and validity of, and training in, ergonomics methods remains an important goal of future research and application. In addition, testing the reliability and validity of new methods during their development is an important requirement in the development of ergonomics methods (e.g. Cornelissen et al, 2015).

References

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

Annett, J. and Stanton, N. A. (2000) Task Analysis. Taylor & Francis: London.

Cornelissen, M., Salmon, P. M., McClure, R., Stanton, N. A. (2014). Validating the Strategies Analysis Diagram: Assessing the reliability and validity of a formative method. Applied Ergonomics. 1484-1494

Crundell, B., Klein, G. and Hoffman, R. R. (2006) Working Minds: A Practitioners Guide to Cognitive Task Analysis. MIT Press: Boston.

Diaper, D. and Stanton, N. A. (2004) The Handbook of Task Analysis for Human Computer Interaction. Lawrence Erlbaum Associates: Mahwah, NJ.

Hollnagel, E. (2003) Handbook of Cognitive Task Design. Lawrence Erlbaum Associates: Mahwah, NJ.

Harvey, C. and Stanton, N.A. (2013) Usability evaluation for in-vehicle systems. Boca Raton, FL: CRC Press.

Nemeth, C. P. (2004) Human Factors Methods for Design: Making Systems Human-Centred. CRC Press: London.

Salmon, P, Stanton, N. A., Gibbon, A, Jenkins, D. and Walker, G. H. (2010) Human Factors Methods and Sports Science: A Practical Guide. CRC Press: London.

Salmon, P. M., Stanton, N. A., Lenné, M., Jenkins, D. P., Rafferty, L. A. and Walker, G. H. (2011) Human Factors Methods and Accident Analysis. Ashgate: Aldershot.

Salvendy, G. (2002) Handbook of Human Factors and Ergonomics, 4th edn, Wiley: New York.

Schraagen, S.F. Chipman, and V.L. Shalin (2000) Cognitive Task Analysis. Lawrence Erlbaum Associates: Mahwah, NJ.

Shepherd, A. (2001). Hierarchical Task Analysis. London: Taylor and Francis.

Stanton, N. A. (2014) Commentary on the paper by Heimrich Kanis entitled 'Reliability and validity of findings in ergonomics research': where is the methodology in ergonomics methods? Theoretical Issues in Ergonomics Science, 15 (1), 55-61.

Stanton, N.A., Baber, C. (2002). Error by design: methods for predicting usability. Design studies, 23, 363-384.

Stanton, N. A. and Baber, C. (2005) Validating Task Analysis For Error Identification: reliability and validity of a human error prediction technique. Ergonomics, 48 (9), 1097-1113.

Stanton, N. A., Hedge, A., Salas, E., Hendrick, H. and Brookhaus, K. (2005a) Handbook of Human Factors and Ergonomics Methods. Taylor & Francis: London.

Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C. and Jenkins, D. (2005b) Human Factors Methods: A Practical Guide for Engineering and Design. Ashgate: Aldershot.

Stanton, N. A., Salmon, P., Harris D., Marshall A., Demagalski J.,Young M.S., Waldmann T. and Dekker S.W.A. (2009). Predicting pilot error: Testing a new methodology and a multi-methods and analysts approach. Applied Ergonomics, 40 (3), 464-471.

Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C. and Jenkins, D. (2013) Human Factors Methods: A Practical Guide for Engineering and Design (second edition). Ashgate: Aldershot.

Stanton, N.A. and Stevenage, S. (1998). Learning to predict human error: issues of reliability, validity and acceptability. Ergonomics, 41, 11, 1737–56.

Stanton, N.A. and Young, M.S. (1998). Is utility in the mind of the beholder? A study of ergonomics methods. Applied Ergonomics, 29, 1, 41–54.

Stanton, N.A. and Young, M.S. (1999a). What price ergonomics? Nature, 399, 197–8.

Stanton, N. A., & Young, M. S. (1999b). A guide to methodology in ergonomics: Designing for human use. London: Taylor & Francis.

Stanton, N. A. and Young, M. S. (2003) Giving ergonomics away? The application of ergonomics methods by novices. Applied Ergonomics, 34, 479-490.

Stanton, N. A., Young, M. S. and Harvey, C. (2014) Guide to Methodology in Ergonomics: Designing for Human Use (second edition). Taylor & Francis: London.

Wilson, J.R. and Corlett, N. (2005). Evaluation of human work, 3rd ed. Boca Raton, FL: CRC Press.