Proceedings of the 18th Conference on Behavior Representation in Modeling and Simulation, Sundance, UT, 31 March - 2 April 2009
[paper: 09-BRIMS-025]

An example of validating models of continuous processes

Dr. Andrew Belyavin, PhD.
Centre for Human Sciences
A50 Building
Cody Technology Park
Ively Road,
Farnborough, Hampshire GU14 0LX
United Kingdom of Great Britain
Telephone: (011)44-1-252 393 401
Facsimile: (011)44-1-252 392 097
E-mail: ajbelyavin@qinetiq.com

Mr. Brad Cain, M.A.Sc., P.Eng.
Defence Research and Development Canada Toronto
Collaborative Performance and Learning Section
1133 Sheppard Avenue West
Toronto, ON, M3M 3B9
Canada
Telephone: 1-416-635-2195
Facsimile: 1-416-635-2013
E-mail: brad.cain@drdc-rddc.gc.ca

Keywords: model, validation, thermal physiology, continuous processes

Abstract: *Validation of models for human behaviour and performance has long been acknowledged as an essential element in the modelling process, but one that is typically conducted poorly if at all. While there may be many reasons that modellers do not validate models, one possibility is that there isn't a formal methodology to provide guidance. This paper will explore a few techniques that may be applied to evaluating models of continuous processes, using the thermal physiological responses of healthy people working in adverse conditions as a demonstration of analyses that could be done. Having done this assessment, we then address the issue of the validity of this model.*

1. Introduction

Validation of human models has been the topic of a number of papers over the past decade since the team headed by Pew and Mavor (1998) published their seminal work on the state of the art of Human Behaviour Representation. Many of these papers lament the lack of validation in Human Behaviour Representation (HBR) and human performance models and while a number do directly compare predictions with observations (e.g. Foyle et al., 2005), many immediately fall back on informal, face validation: TLAR (that looks about right[1]) or BOGSAT (bunch of guys sitting around the table: Campbell & Bolton, 2005). To be fair, the number of papers that do compare predictions to observations with formal analytical methods (e.g. Gluck & Pew, 2005) is growing and perhaps what the community needs is a broad corpus of worked problems to establish that formal validation is not only an essential step but also a feasible step in model development.

For many, colloquial definition of the validity of a concept or a model means accurate representation of real world events (Trochim & Donnelly, 2007) although there have been numerous attempts to describe just what is meant by validation in a scientific context that are more useful (Anastasi, 1997; Cronbach & Meehl, 1955). Absolute comparisons with the real world may not be the most appropriate starting point for addressing the validity of a model. Formal models are typically abstractions of the processes that we believe explain the observed events to some degree, but models often ignore aspects of the real world experience, either through ignorance or design. Thus, trying to validate a model as an accurate representation of the real world events is, in this strict sense, doomed to failure.

Perhaps a more appropriate strategy is to acknowledge that validation means "fitness for purpose". George Box (1979) is often quoted to have observed that "all models are wrong, but some models are useful." This statement captures the essential aspect that we feel

---

1 http://acronyms.thefreedictionary.com/That+Looks+About+Right

should be the topic of validation: Does the model capture and display enough of the real world observations to produce a result that is useful for the current application?

1.1 Philosophical Issues

Discussions of how best to validate models have been going on for some time and the answer would seem to rest on similar concerns to that of theory validation, if one considers a formal, computational model to be the instantiation of a theory (whether or not it is explicitly expressed as such.) The debate about what is a suitable method of validation seems to revolve around two extremes: Null Hypothesis Significance Testing (NHST) versus Goodness of Fit (GOF). The interested reader can find published uses of each, often decrying the use of the methods favoured by the other camp (see the following series of articles for such a discussion: Roberts & Pashler, 2000, 2002; Robinson, Duursma & Marshall, 2005; Rodgers & Rowe, 2002).

In practice, we suspect that the best approach will lie in the middle ground, involving a mixture of methods from both camps, and indeed, that seems to be where the above mentioned discussion seems to end. In this paper we have taken a currently contentious position for assessing validity, that of starting with NHST methods. Obviously, a model must make predictions that at least resemble the observations before it is evaluated, but from our perspective, it seems that determining model validity (its fitness for purpose) and determining how well the model captures reality (as noted above), it is more important to determine the magnitude and the patterns of differences of model predictions compared with observations rather than determining their degree of agreement.

The choice of technique assessing the differences between model and observation is a matter of what hypothesis is being tested. When it is simple effects of the context on model outputs, it may be appropriate to analyse some sort of model and observed means and use a version of Analysis of Variance (ANOVA). When it is some sort of random effect, partially explained by the model, it appears more appropriate to use regression or GOF.

We contend that a good approach to the analysis is the use ANOVA initially to identify how the differences partition over the various effects. We do not see the utility of single goodness of fit metric for determining these differences. For a complex multi-component model of this kind an initial ANOVA provides an indication of how the differences break down over the different pieces. More detailed analysis is required to identify where the model appears to work and where it does not, and these subsequent steps may well include GOF methods.

This paper will discuss a few techniques that can be used to assess models; in particular, models of continuous processes that are sampled periodically. We have selected a rational model of thermal physiology as an exemplar and we will discus its use within the context of a physiological experiment involving subjects working in a hot environment using found data. Although this application is somewhat removed from the harder problem of assessing behaviour, we feel that validating models of operator state is nevertheless an important element in characterizing individual differences in Human Behaviour Representation and will hopefully provide a starting point for further exploration of validation methodology.

2. Validation Approach

2.1 Empirical Experimental Method

Data was obtained from a within-subject physiological experiment (Cheung & McLellan, 1998)[2], conducted with 8 healthy male volunteers exercising in a chamber at 40°C and 30% relative humidity while wearing the 1990s version of the Canadian Forces NBC[3] protective garment (thermal insulation: 0.29 m^2K/W or 1.88 CLO; water vapour diffusion resistance: $4x10^7$ m^2 Pa s/kg or 0.33 $I_{m,}$ Woodcock Permeability Index, as determined by measurements with wetted thermal manikins, sweating hot plates and vapour permeability devices).

Subjects walked on a treadmill in two conditions: a low workload condition walking at 3.5 km/h with no grade, approximately equivalent to a total metabolic rate of approximately 165 W/m^2; a high workload condition walking at 4.8 km/h and 4% grade, approximately equivalent to approximately 260 W/m^2.) Subjects began the sessions either at normal hydration levels (euhydrated) or dehydrated by approximately 2% (hypohydrated). Additionally, in the euhydrated contrition, subjects either had no water or drank 200-250 ml of 37°C water every 15 minutes (to offset sweating) while in the dehydrated condition, subjects were given water as above.

The subjects' weight, height, percent body fat and VO2max were measured in a session prior to experimental testing. A break of approximately 1 week was scheduled between each experimental condition to minimize order affects. Subject rectal and weighted skin temperatures and metabolic rate were recorded every 5 minutes; total sweat and evaporation rates where calculated based on measurements of pre- and

---

[2] Defence and Civil Institute of Environmental Medicine protocol L127

[3] NBC: Nuclear, Biological, Chemical

post-exercise body weights, including any water consumed during each trial.

2.2 Analytical Experimental Method

A simple IPME (Integrated Performance Modelling Environment) task network was used to set up environmental conditions, initialize individual subject characteristics (intended to predict individual differences), stimulate the thermal physiology model and record the results. A rational, multi-segment, 1-dimensional thermal physiological model was used in our evaluation, representing the physiological "controllers" and flow of heat and fluid (Higenbottam & Belyavin, 1998). A published exercise model (Givoni & Goldman, 1971) was used to predict the metabolic rate of walking, the exercise undertaken by the subjects in the empirical study.

The model operators exercised at the same rates and under the same conditions as the subjects as specified in the empirical protocol but because the thermal model is not sensitive to hydration status, only a portion of the empirical study was replicated for this paper: conditions where subjects were not dehydrated and were given water during the trials to offset sweating. The model was sensitive to environmental conditions, clothing characteristics, activity level and individual anthropometric differences. Predicted responses for the rectal and skin temperatures were recorded every 5 simulated minutes, to correspond to the experimental observations.

2.3 Comparison Approach

There are two elements to the process of making a comparison between observations drawn from human experimentation and modeled results with a view to determining model validity: understanding the pattern of differences between the two sets of observations and determining whether the criterion of acceptability for model validity for the purpose in mind has been met.

There have been attempts to set up statistical tests and criteria that characterise acceptable limits of agreement among data drawn from models and experiments (Grant, 1962; Robinson et al., 2005). These tests were conceived to avoid the embarrassing situation of attempting to prove the null hypothesis (H0: model and observation data are from the same population) as evidence that model and observation are indistinguishable. Further compounding the issue is the recognition that a weak design with low power may indicate no significant difference while a more powerful study may lead to accepting the alternative hypothesis (H1: model and observation data are from different populations) even though the differences could conceivably be of little practical significance (Campbell & Bolton, 2005).

The situation of validating a model is no different from that of assessing a scientific theory. At no stage can we "accept" the null hypothesis that model and reality conform adequately as a permanent solution; any "acceptance" is strictly provisional. If the data do not provide a sufficiently powerful test of the null hypothesis, the solution is to collect more data, not to argue that we are caught in a paradox. In addition, science has a long history of successfully falsifying theories such as Newton's laws of motion but also demonstrating that the theory can be applied in a restricted set of circumstances.

It should be noted that statistical methods, when properly applied, represent tools that can provide analysts and researchers with insight into analysis of data. The application of statistical methods never "proves" anything in a formal sense, neither the null nor the alternative hypotheses, but they provide evidence upon which a provisional judgement may be made. Having a summary statistic that assesses the degree of agreement between model and observation **may** be useful to express how well a model fits the data, but it fails to address the issues of "What are the differences?" and "Are the differences important to the intended application?"

We make the assumption that an empirical study conducted by subject matter experts in a field (in this case thermal and exercise physiology) that produces reliable results that distinguish between experimental conditions would also serve as a reliable assessment of the performance of a model. Our hypothesis is that if the model is a valid representation of the thermal physiological responses to the experimental conditions, then not only will it predict the mean, normative responses but it will also reflect the observed variance attributable to individual anthropometric differences in a rational manner. Of course any generalizations based on these conclusions of validity should be limited to similar operational conditions for the relevant factors.

We do not assume that differences between model and observation **necessarily** render the model invalid; nor do we assume that a null result (no significant difference) renders the model valid. This approach to validation is based on a whole-hearted acceptance of the proposition that no model will capture all aspects of reality to perfection and it is the manner and degree of mismatch that is important for determining the utility of the model.

It is reasonable to express this concept in an acceptable level of inconsistency between model and real world observation that can be tested in the spirit of Robinson et al. (2005) GOF approach. Suitable metrics might be mean absolute difference between model and observation, or proportional difference expressed by

the slope of a regression line. Difficulty arises if the pattern of differences observed in the model does not conform to the prior model; for example the difference turns out to be proportional when a mean difference was regarded as an acceptable criterion. We argue that such differences should be analysed and their characteristics established, before final determination as to whether they fall within an acceptable region. In addition, for complex models, as in the case of a whole body thermal model, it seems wise to break down the analysis into the component models before drawing a final conclusion about the ensemble, as performance of the whole model may be unacceptable despite some of the component models behaving adequately or *vice versa*.

In the case chosen to illustrate our approach the observed data are drawn from a designed experiment that would be investigated using standard statistical methods such as ANOVA. When ANOVA is applied to a designed experiment it is first necessary to determine which of the possible ANOVA models should be applied in terms of the assumptions that are made about the nature of the controlled factors. In the case of the data under consideration there are three controlled, independent factors: external work level, measurement period through the exposure and the subject. The first factor has two levels – low work and high work characterized by the grade and speed of walking imposed by the treadmill – and should be treated as a fixed factor. The second factor – measurement period through the exposure – has a nominally fixed set of levels and is also a fixed factor. The third factor, the subject, is conventionally treated as a random factor – the levels of which represent different sampled elements from an assumed infinite population. The set of participants are all exposed to the same combination of fixed factors so there are repeated measures made on the random factor.

The key statistical assumptions of the repeated measures model are threefold:

- The random factor is sampled from a normally distributed population;
- The errors of observation are independently sampled from a common normal population;
- The observation errors and the random effects are additive and independent.

These assumptions can be expressed in a different form in that if the repeated observations are considered as a multivariate vector, the observations can be shown to be drawn from a multivariate normal population with fixed mean and diagonal covariance matrix with common variances – the sphericity assumption. One of the fixed factors in the experiment – progress through the exposure – is particularly vulnerable to violation of the underlying assumptions as there are likely to be autocorrelations between successive observations. A further complication is that experimental runs were terminated when participants' temperatures exceeded pre-defined thresholds or when subjects elected to withdraw from a trial so that there are a different number of observations for each subject in each condition. These missing observations are dependent on the experimental conditions and should be modeled in a more complex manner, most plausibly as censored data. For convenience, we have restricted the comparison between model and observations in this report to the observations made over the shortest duration trial by a subject in each condition, ignoring data for other subjects after this time, so that more complex assumptions are not required.

The thermal model under test has been designed to capture individual subject variability attributable to anthropometric differences. As we are interested in modelling both the expected value and variability of physiological responses, the subjects' physiological characteristics body weight and percent body fat were entered into the model, creating a pseudo within-subject and repeated-measures design. Although the empirical observations and analytical model predictions are not truly a repeated-measure, the thermal model incorporates individual differences in an attempt to accomplish this goal and was thus treated as a further fixed effect, subject exposure repetition in the analysis.

When validating human performance models there are two issues: are the gross changes caused by external conditions modelled appropriately and are the random variations due to individual differences captured. In our example for thermal physiology, the latter comprise two pieces: physical, anthropometric characteristics and possible variations in the various phenomena controllers embedded in the model (e.g. metabolic rate, blood flow, sweat rate, etc.). To identify the effects of the distinct components it is necessary to apportion the potential match between model and observations over these different sources of variation rather than use a gross goodness-of-fit. The ideal vehicle for this dissection is the ANOVA family of statistical methods.

Three primary measures were considered in the analysis for the current report: rectal temperature, mean skin temperature and metabolic rate. These variables were selected to demonstrate the techniques used rather than as a complete assessment of the accuracy of the model, which would include a comprehensive assessment of all the pertinent variables.

3. Results

There are multiple models that comprise the prediction of thermal behaviour and three key components can be identified: exercise model, clothing model and internal whole body model. The particular trial used for validation purposes in this paper is primarily dominated by the imbalance between the heat generated by exercise and the heat lost by the evaporation of sweat through the clothing. It is possible to assess the exercise model as a distinct component by comparing observed and predicted metabolic rates. It is not possible to assess the clothing model as no measure of sweat evaporation is available with the current empirical data set.

3.1 Metabolic rate

Observed and predicted metabolic rates were compared in an ANOVA to assess whether the means were different between observed and predicted values. From the analysis it was concluded that there is a difference in metabolic rates between High and Low Work conditions ($F=479.6$, $df=1,7$, $p<0.001$), but it was not possible to demonstrate a difference in the mean between model and observations in general ($F=2.97$, $df=1,7$, $p>0.05$) or an interaction between model and observation for the two work levels ($F=3.32$, $df=1,7$,$p>0.05$). The means are displayed in Figure 1.

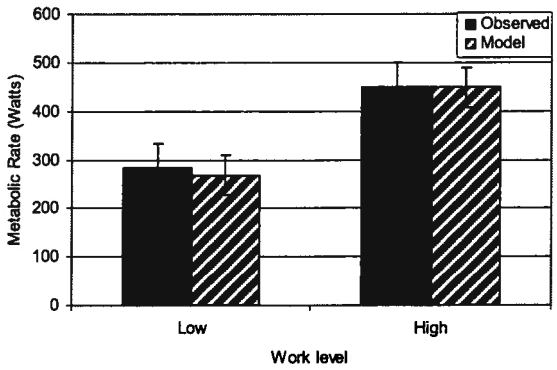

Figure 1. Comparison of observed and predicted mean metabolic rates with 95% confidence intervals.

There is substantial variability between the subjects in the observed data that is also demonstrated in the modelled results. The relationship between the two measures for both work conditions is displayed in Figure 2. To test whether the variability is captured by the model an Analysis of Covariance (ANCOVA) was conducted using the modelled value as the covariate and the observed value as the dependent variable. It was found that both the variability between different participants and the difference between Low and High Work conditions in the observations could be explained

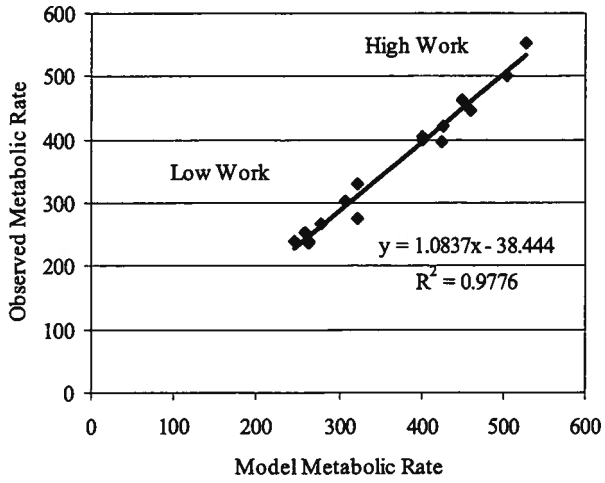

Figure 2. Comparison of Observed and Predicted Metabolic rates for 8 subjects under 2 work loads.

by the model indicating a high level of accuracy for the exercise model and an expectation that it would be valid for many similar conditions. The use of ANCOVA in this case is similar to a goodness of fit.

3.2 Rectal temperature

Observed and predicted rectal temperatures were compared separately for High and Low work conditions. To restrict the possibility of autocorrelation between successive errors, remove any differences in assumed initial conditions and minimise lost data, the comparison was restricted to 7 observation times (Period): 15, 30, 45, 60, 75, 90 and 105 minutes after the start of the exercise period in the Low Work condition. The means across subjects are displayed in Figure 3 for each Work condition.

An ANOVA was conducted to assess mean differences between observed and modelled values at the 7 time Periods with Modelled/Observed as fixed factors and Participant as a random effect. It was concluded that there is a main effect between the Periods ($F=540.23$, $df=6,42$, $p<0.001$) and an interaction between Modelled and Observed with Period ($F=7.72$, $df=6,37$, $p<0.001$) although there was no evidence to support an overall difference, or main effect, between modeled and observed values ($F=0.0$, $df=1,7$, $p>0.05$).

A similar analysis was conducted for the High Work case for 4 times: 15, 30, 45 and 60 minutes. It was concluded that there is a main effect of Period ($F=1114.46$, $df=3,21$, $p<0.001$) and an interaction between Modelled and Observed for the different Periods ($F=14.95$, $df=3,16$, $p<0.001$). The difference between Modelled and Observed values was not found to be statistically significant ($F=2.71$, $df=1,7$, $p>0.05$).

It is clear from Figure 3b that the model is predicting a faster rise in rectal temperature than is observed in the experiment since Modeled values exceed Observed values at Periods 3 and 4.

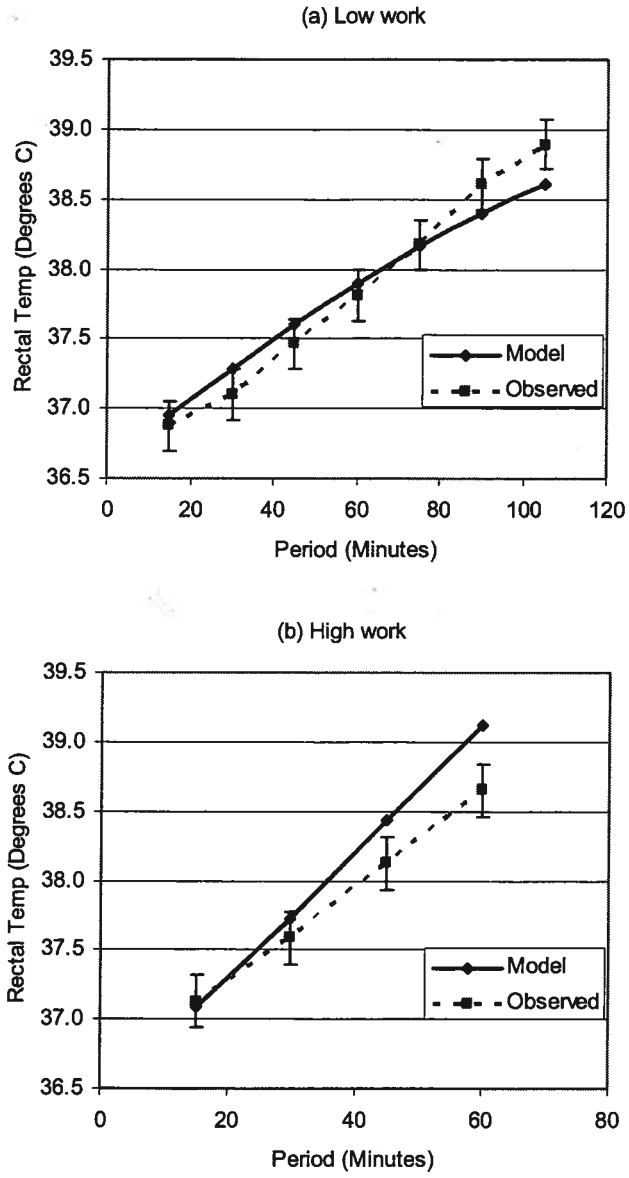

Figure 3. Modelled and observed mean rectal temperatures with 95% confidence intervals for the two levels of work intensity.

There is clear variability between the subjects' rectal temperatures in response to exercise. A simple, main-effect comparison between the observed and modelled means (averaged over the periods) was examined and the findings are displayed in Figure 4. As there was no association between observed and modelled participant means, it was concluded that the model does not capture the random individual, variability in rectal temperature, in contrast to the exercise model.

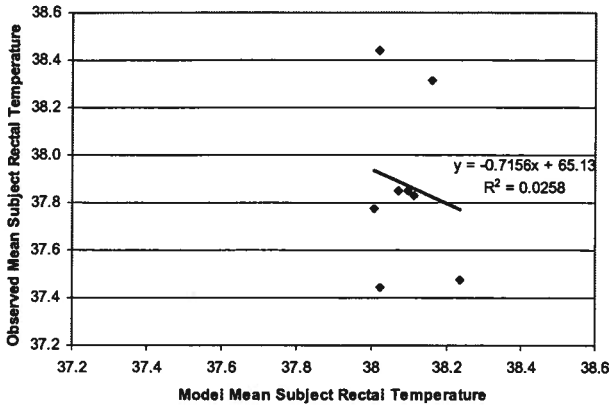

Figure 4. Observed vs Modelled Average Subject Rectal Temperatures during High Work condition.

A plot of the observed and corresponding predicted rectal temperatures is shown in Figure 5. including a linear regression over all these data. Although this regression is not as elegant as other techniques in the literature, it is included here as an example of a simple GOF metric for illustration of a point. As shown, the correlation coefficient seems reasonably high, with the model capturing almost 85% of the observed variance. Yet the ANOVA analysis indicates that a decision of validity based on this measure that we are capturing the variability due to subject differences is unjustified.

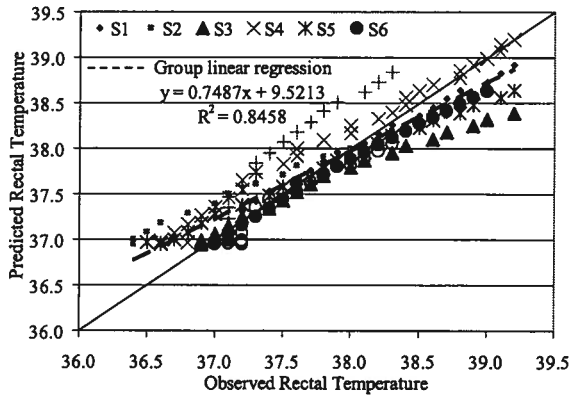

Figure 5. Plot of observed and predicted rectal temperatures by subject for the low work condition used in the analysis with a linear regression fit.

3.3 Skin Temperature

Similar analyses were conducted for the skin temperatures as were performed on the rectal temperatures and mean values are displayed in Figure 6 for each condition. In the Low Work condition there is a main effect of Period ($F=313.13$, $df=6,42$, $p<0.001$) and a main effect of Model versus Observed values

(F=8.72, df=1,7,p<0.05) but there is an interaction between Modeled and Observed values with Period (F=6.15, df=6,37, p<0.001).

In the High Work condition, main effects of Period (F=629.722, df=3,21, p<0.001) and Modelled versus Observed (F=7.112, df=1,7, p<0.05) were found but again there is a significant interaction between Modelled and Observed values with Period (F=40.61, df=6,37, p<0.001).

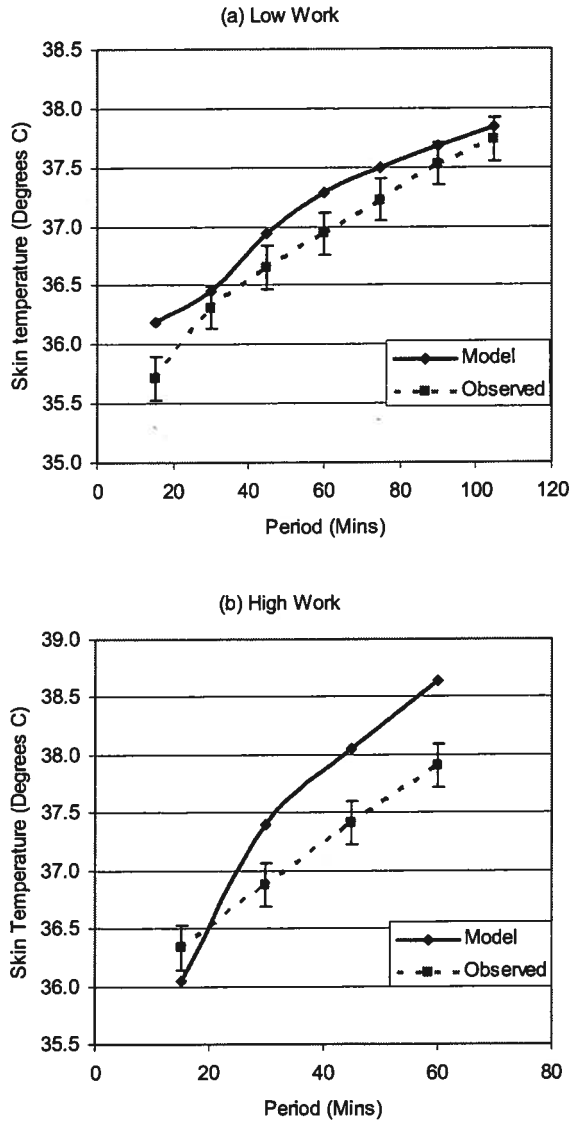

Figure 6. Modelled and observed mean skin temperatures with 95% confidence interval for the two levels of work intensity.

4. Discussion

The characterisation of the discrepancy between model results and those drawn from a single experiment indicate that the model predictions of metabolic rate as a consequence of exercise cannot be distinguished from the set of real world observations. On the data available to test this model component, there is no reason to reject its use for the current purpose. This does not mean that this model is unconditionally validated, merely that there is no reason at the moment to reject its future use based on this analysis. We do not make this judgement based on the failure to reject the null hypothesis in the comparison, but we use that failure as a starting point for determining if the agreement is sufficient for our purposes. This constitutes the gold standard in predictive validity and the model is clearly potentially useful as a predictor of outcomes in the world.

The characterisation of the pattern of differences between observed and modeled values of rectal and skin temperatures indicates that the model has not predicted the response of rectal and skin temperatures change with exercise with the same level of accuracy as the metabolic rate model. In addition, at least for rectal temperature, the model has not reproduced the pattern of variability either within or between participants in the observed data.

The question remains as to whether the model is useful. It is clearly possible to adopt the posture that the model does not predict the pattern of observations and displays significant differences from observation; it is therefore not exactly correct and cannot be safely used for any purpose. We argue that this is an extreme position and it is not an appropriate interpretation of the analysis in a fitness-for-purpose assessment. It is clear to us that the model is not projecting the observed variability in rectal temperatures to any degree we conclude that the model is not useful for predicting results where individual variability is important to its intended application, and thus this model is not valid for that purpose.

The question remains as to whether the model can be used for predicting population mean responses in a useful way. From the analysis it can be concluded that the model is not predicting the way that rectal temperature increases in either condition. The extreme mean error in the High Work condition is 0.48°C and in the Low Work condition is 0.29°C. These represent approximately 30% and 20% of the rise in rectal temperature in each case, and it is known that the discrepancy is positive in the High Work case and negative in the Low Work case. Confidence intervals can be constructed for these measures on the basis of the variability of the observed data, taking account of the variability between participants or otherwise. It is likely that for the range of exercise that lies between these two values the error is likely to fall between these

bounds. *If these bounds are acceptable to the end user, the model as applied in this case is useful.*

We argue that the job of the analyst is not to define some goodness of fit criterion and use this statistic to decide validity as it very likely will obscure much of the detail of how model and observation differ. Instead, we feel that the analyst should clearly present the pattern of discrepancy between model and observation and then argue whether the discrepancy is sufficient to render the model useless on the basis of how the model will be applied. The same reasoning applies to the analysis of the applicability of Newton's laws of motion.

It can be argued that a single set of data that may or may not indicate discrepancies between model predictions and observations does not provide sufficient basis for describing model behaviour. In this particular case there is not a clear conclusion that the model over-predicts the rise in core temperature, suggesting that a potentially complex pattern of discrepancy may underlie what has been observed. This argument is valid regardless of the outcome of the analysis and whether the model matches observations or not should not determine whether further data sets are investigated. The key to successful application of a model will always be based on an understanding of how the model behaves across the range of conditions in which it could be used. The largest number of validating data sets should be used to describe model performance across the range of conditions, particularly if an emergent property of the model demonstrates uncommon, yet plausible results while excluding predictions of implausible responses.

As well as providing evidence for model performance, the approach we've followed provides information that may be used to improve the current model's performance in future. The existing model is not representing the pattern of heat loss from the body and through the clothing appropriately, in that both core and skin temperature increase more than the observations. It might be possible to adjust the model by amending the representation of sweating, clothing or environmental conditions to conform to the current observations and then to revalidate the model with further data, preferably data that span distinctly different conditions.

5. References

Anastasi, A. (1997). Validity: Basic concepts. In A. Anastasi & S. Urbina (Eds.), *Psychological Testing* (7 ed., pp. 113-139). New Jersey: Prentice Hall.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201-236). New York: Academic Press.

Campbell, G. E. & Bolton, A. E. (2005). HBR Validation: Integrating Lessons Learned From Multiple Academic Disciplines, Applied Communities, and the AMBR Project. In K. A. Gluck & R. W. Pew (Eds.), *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation* (pp. 365-395). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cheung, S. S. & McLellan, T. M. (1998). Influence of hydration status and fluid replacement on heat tolerance while wearing NBC protective clothing. *European Journal of Applied Physiology, 77*, 139-148.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Foyle, D. C., Hooey, B. L., Byrne, M. D., Corker, K. M., Deutsch, S., Lebiere, C., et al. (2005). *Human performance models of pilot behaviour.* Paper presented at the Human Factors and Ergonomics Society 49th Annual Meeting, Santa Monica, CA. 1109-1113.

Givoni, R. & Goldman, R. F. (1971). Predicting metabolic energy cost. *Journal of Applied Physiology, 30*, 429-433.

Gluck, K. A. & Pew, R. W. (Eds.). (2005). *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69*(1), 54-61.

Higenbottam, C. & Belyavin, A. J. (1998). *2D Thermoregulatory Model - Description and Validation* (No. DERA/CHS/PPD/WP980204/1.0). Farnborough, Hampshire: Defence Evaluation and Research Agency.

Pew, R. W. & Mavor, A. S. (Eds.). (1998). *Modeling Human and Organizational Behavior. Applications to military simulations. Panel on Modeling Human Behavior and Command Decision Making: Representations for Military Simulations. Commission on Behavioral and Social Sciences and Education.* Washington, DC.: National Research Council. National Academy Press.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Roberts, S. & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review, 109*(3), 605-607.

Robinson, A. P., Duursma, R. A. & Marshall, J. D. (2005). A regression-based equivalence test for model validation: shifting the burden of proof. *Tree Physiology, 25*, 903-913.

Rodgers, J. L. & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review, 109*(3), 599-604.

Trochim, W. M. K. & Donnelly, J. P. (2007). *The research methods knowledge base* (3 ed.). Mason, OH, USA: Thomson.

Author Biographies

Dr. Andy Belyavin is a QinetiQ Fellow specializing in statistical analysis and human performance modeling, and is a leading technical consultant on the application of statistical methods to the analysis of data ranging from experimental results to complex pattern recognition. He graduated BA in Mathematics at the University of Cambridge, and PhD in Applied Statistics at the University of Reading. For the past 10 years, he has led the development of the Integrated Performance Modelling Environment (IPME) under contract to the Ministry of Defence (MOD). Within this project, he is working on human performance for IPME, including quantitative models of the effects of various stressors upon performance, further development of workload models in conjunction with DRDC Toronto, and models of human tracking performance. In addition, his work includes the development of manpower models, and he acts as a source of advice and guidance on the development of applications involving both physiological and psychological models, and has been associated with the development of some models and statistical techniques to support the QinetiQ augmented cognition program.

Brad Cain is a Defence Scientist at Defence Research and Development Canada Toronto. His background is originally in modeling heat transfer and fluid flow, developing personal protective clothing and equipment for the Canadian Forces. For the past several years he has been developing models of human performance and tools to support performance and cognitive workload modeling for applications in simulation-based acquisition and distributed training systems. He is the Canadian lead on the collaborative development of the human modeling software IPME, working with QinetiQ Ltd and Alion Science and Technology. Currently, Brad's main focus is a program to support human behavioural modeling approaches that can serve as virtual agents in distributed simulations that build on human engineering modeling tools and approaches, yet incorporate human sciences knowledge. Brad is a Professional Engineer, graduating from the University of Toronto in Mechanical Engineering (BASc 1979, MASc 1981).