Multi-model Approach to Human Functional State Estimation

Kevin Durkee[1(✉)], Avinash Hiriyanna[1], Scott Pappada[1], John Feeney[1], and Scott Galster[2]

[1] Aptima, Inc., Fairborn, USA
{kdurkee,ahiriyanna,spappada,jfeeney}@aptima.com
[2] Air Force Research Laboratory, Dayton, USA
scott.galster@us.af.mil

Abstract. With the growth and affordability of the wearable sensors market, there is increasing interest in leveraging physiological signals to measure human functional states. However, the desire to produce a reliable universal classifier of functional state assessment has proved to be elusive. In efforts to improve accuracy, we theorize the fusion of multiple models into a single estimate of human functional state could outperform a single model operating in isolation. In this paper, we explore the feasibility of this concept using a workload model development effort conducted for an Unmanned Aircraft System (UAS) task environment at the Air Force Research Laboratory (AFRL). Real-time workload classifiers were trained with single-model and multi-model approaches using physiological data inputs paired with and without contextual data inputs. Following the evaluation of each classifier using two model evaluation metrics, we conclude that a multi-model approach greatly improved the ability to reliably measure real-time cognitive workload in our UAS operations test case.

Keywords: Context · Human performance · Modeling and simulation · Physiological measurement · Workload · UAS · Cognitive states

1 Introduction

With the dramatic growth and affordability of the wearable sensors market in recent years, there is increasing interest throughout many work domains in leveraging human users' real-time physiological signals to measure functional states, such as workload, stress, and fatigue. In military defensive settings, the ability to monitor these states throughout a mission would be a valuable asset to optimize mission operations and warfighter workflow. As the complexity of military operations continues to increase, warfighters will become increasingly vulnerable to undesired cognitive states. Measuring cognitive states in relation to task and mission performance would provide the requisite data to detect if, and when, a warfighter has met his/her limits while diagnosing what intervention is best suited to sustaining good performance and obtaining the desired outcomes. By introducing this capability, assessments of operator states would become

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2016, Part I, LNAI 9743, pp. 188–197, 2016.
DOI: 10.1007/978-3-319-39955-3_18

integral system parameters about the mission to be proactively monitored and addressed before potential problems occur [1].

The ability to obtain real-time physiological data carries promise for providing such a capability to the warfighter. Physiological data have the substantial benefit of being an objective source of information that is theoretically available from any person working in any domain. The use of physiological data to classify a human operator's state has been extensively researched over the past few decades, frequently suggesting the existence of measurable indicators that are predictive of a particular state, or change in state [2, 3]. The majority of research has employed some combination of electroencephalography (EEG) [4], electrocardiography (ECG) [5], pupillometry [6], or galvanic skin response (GSR) [7]. For example, in the Air Force Multi Attribute Task Battery (AF_MATB) environment, Wilson and Russell (2003) introduced a novel application of artificial neural networks (ANNs) trained to each individual human performer for real-time mental workload classification using six channels of brain electrical activity, as well as eye, heart, and respiratory signals [2]. Wang et al. (2012) also employed the AF_MATB to introduce a novel hierarchical Bayesian technique that showed promise for cross-subject workload classification [3]. Although the majority of these and other studies have been laboratory based and often employ costly and/or invasive monitoring equipment, recent improvements in sensor reliability, level of invasiveness, set-up time, and cost have made the concept more compelling for high-fidelity work environments.

Assuming sensor limitations are eventually overcome, as current trends would suggest, several additional limiting factors still exist that have hindered progress in the area of human functional state assessment. Perhaps the most notable challenge is lack of consistent physiological indicators of a particular state or change in state. Most commonly these inconsistent patterns in physiological signals occur as a function of individual differences across people. However, this issue can also frequently occur across time (e.g., different days or times of day) for a specific individual person. These differences can be drastic; for instance, a highly reliable indicator of workload level for one person could provide no utility in assessing workload for another person, or vice versa. Prior research has often been forced to cater to this challenge by training a classifier on a per-person, per-day basis [2], or by including a given person's data in both the model training and model testing sets [3], neither of which is practical for implementing in real-world environments. Additionally, some individuals have few, or weak, discernible physiological indicators of a functional state, making it difficult to build a reliable model to classify a state of interest.

Over the past few years, we have made progress designing a universal machine learning based approach to pinpoint a human operator's state with high resolution (0-100 scale) and update frequency (second-by-second) with physiological-based assessment [8]. Our concept was further expanded to evaluate the added precision offered by integrating contextual data with physiological signals within a Functional State Estimation Engine (FuSE^2) [9]. The addition of contextual data in particular was shown to provide noteworthy improvements to the challenge of inconsistent and/or weak patterns in one's physiological signals. Although this comes at the cost of a model classifier being tied to a specific task environment, these results did not require the use of personalized models that were trained to a specific individual [9].

In spite of these improvements, the desire to produce a universal computational model for functional state assessment has proved to be difficult, and there remains significant room for innovation to solve this problem. For this reason, we have continued to investigate novel and supplementary strategies for more consistent model classifier results that would provide the necessary reliability for real-world utilization. One concept that has not been thoroughly explored in human functional state assessment is the convergence of multiple model classifiers into a single measurement of state. There is evidence in other related fields such as adaptive system development that basing decisions on a multi-model approach can outperform the same decisions being made from a single-model approach [10]. We theorize a similar approach would improve human state assessment given the many different ways that one's physiological data could be modeled, each having its own unique benefits and drawbacks. In addition, for situations in which no given model is able to accurately measure a human functional state, a multi-model approach can increase our confidence that the measurement challenges may lie in the data set itself (e.g., due to lack of distinguishable patterns), rather than a flaw in the use of machine learning and model development. This would necessarily shift the focus toward the need for more distinct and consistent sources of sensor data that can better indicate a person's functional state.

The objective of this work was to explore to what extent a multi-model approach can increase the accuracy of physiological-based cognitive state classifiers in a UAS task environment. In particular, our goal was to explore the multi-model approach from a bottom-up perspective by decomposing a cognitive state of interest into multiple subcomponents that are each individually modeled and subsequently fused together to build the construct. In the following sections we review a UAS study that was used to produce data for building real-time workload classifiers within the FuSE^2 system using both the single-model and multi-model approaches for comparative analysis. We also opted to examine the effects of adding two contextual data inputs– human computer interaction (HCI) rate and primary task performance – to investigate if the effect of using a multi-model approach remained present after the accuracy boost presented in our previous analysis [9]. Although FuSE^2 is capable of on-line supervised learning to adapt to an individual for improving model accuracy, we restrict the scope of this paper solely to cross-subject workload classification since a universal "plug and play" model that does not require per-subject training would be an ideal technological milestone.

2 Methods

2.1 Data Collection

Data were collected within a simulated UAS task environment – the Vigilant Spirit Control Station (VSCS) [11] – at the Human Universal Measurement and Assessment Network (HUMAN) Laboratory located at Wright-Patterson Air Force Base. We focused exclusively on cognitive workload for this study and the ensuing model development effort so as to constrain the problem space to a single human functional state that has wide applicability, particularly to UAS operations, and a large body of literature to draw from as needed. The UAS task simulation employed the VSCS operator interface

(Fig. 1) paired with a Multi-Modal Communication (MMC) tool for issuing communication requests [12] and a custom-built lights and gauges monitoring display. The primary task objective was to track a high value target (HVT) while keeping the HVT continuously positioned on the center of the UAS sensor crosshairs. Simultaneously, participants conducted two secondary tasks: (1) monitor the lights/gauges display and acknowledge each system event via button presses; and (2) verbally respond to each communication request via the MMC tool. Task difficulty was manipulated by modifying the HVT speed and motion complexity, the number of communication requests, and the number of light/gauge events in each five-minute trial. This task paradigm allowed for a gradual titration of task difficulty across 15 five-minute conditions ranging from easy to hard, which was intended to induce variations in workload and performance for each participant.

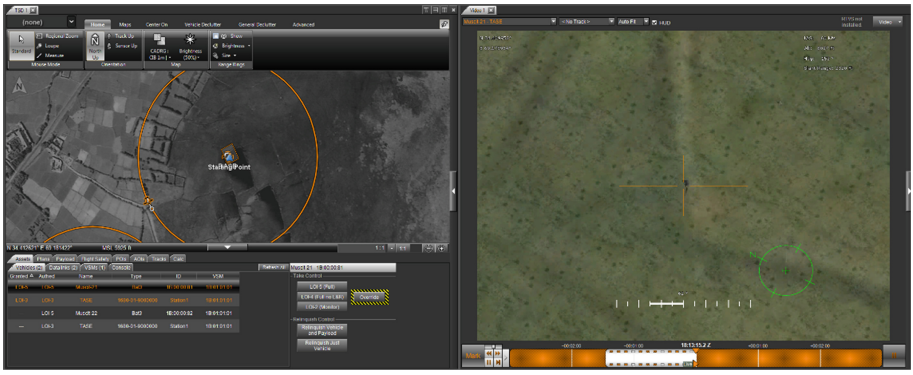

Fig. 1. The VSCS operator interface

There were 25 participants with each person completing one training session and one data collection session each. Dependent measures were threefold: (1) a suite of physiological metrics collected during each task condition consisting of six-channel EEG, ECG, off-body eye tracking, respiratory activity, electrodermal activity, and voice analysis features; (2) self-reported NASA Task Load Index (TLX) responses collected at the end of each trial [13]; and (3) system-based performance measures derived from Aptima, Inc.'s Performance Measurement Engine (PM Engine^{TM}) that utilized behavioral and situational data to estimate continuous performance for all three task requirements. NASA TLX responses and condition difficulties yielded a correlation of $r = 0.75$ across all subjects and $r = 0.89$ mean correlation within subjects, suggesting the manipulations were successful at inducing the intended variance in workload.

2.2 Model Development

Using the data collected from this study, a set of model-based classifiers was developed within the FuSE^2 system using machine learning techniques that train each classifier to output second-by-second workload estimates on a 0–100 scale. In accordance with the goals of this analysis, classifiers were trained for both the single-model and multi-model

approach. The single-model approach was implemented in which the model inputs were trained directly to the composite NASA TLX response of total workload. In contrast, the multi-model approach was implemented in which a set of model inputs was trained to each of the six sub-scales of the NASA TLX provided by study participants. After the NASA TLX sub-scale models were trained as part of the multi-model approach, participants' sub-scale card sort data were used to weight each model's respective contribution in determining total workload. For both the single-model and multi-model approaches, this model training approach was done twice: once without contextual data as a model input, and once with contextual data as a model input.

Development of each model classifier adhered to the approach in Durkee et al. [8], in which we applied a noise injection algorithm to all NASA TLX responses under the assumption that workload does not remain perfectly static over time. This algorithm derives an estimate of "ground truth" on a second-by-second basis to which the model classifiers are subsequently trained. We refer to each series of ground truth estimates as the "desired model output" given each model classifier's attempt to find the best fit based on its feature inputs. Because it is impractical to obtain operator responses at very frequent intervals, this algorithm relies on a theoretically-grounded correlate of workload as the basis for injecting this noise. Although the same correlate was used for noise injection in the single-model approach as was done in our prior work [8, 9], the selected correlates for the sub-scale models in the multi-model approach varied. This was done because the six sub-scales that produce a NASA TLX value each have unique innate qualities that vary in different ways (e.g., mental demand varies based on mental activity, whereas physical demand varies based on physical activity). Hence, it was assumed the same correlate should not be used across these sub-scale models, and as such, careful consideration was given based on scientific theory supported with empirical literature.

Following the noise injection stage to produce desired model output values for training, a comprehensive training set was prepared containing all selected feature inputs and the desired model outputs. The training set included data from 19 of the 25 study participants, while the other six participants were randomly selected for model evaluation. A training process was initiated to derive model weights for each classifier based on minimizing error between the feature inputs and the desired model outputs. The selected physiological inputs for all model classifiers were three EEG channels (Fz, Pz, O2), ECG, and pupillometry; and, as previously mentioned, two versions of all models were created: one without context, and one with context.

2.3 Model Evaluation

After completing the model training process, the next objective was to produce test results in order to evaluate the accuracy of each workload classifier, particularly to assess how the multi-model approach impacted model accuracy relative to the single-model approach. Workload classifier results were produced through a batch playback of data collected from the six participants excluded from the training set. All six test participants completed the same 15 five-minute trials used to train the model classifiers, thus totaling 90 trials used for evaluation. The batch playback process simulated the production of

real-time classifier results by outputting one workload estimate per second on a 0–100 scale for all models, totaling 300 values per model within each trial.

Model accuracy was analyzed via summary statistics in two ways: correlation and absolute difference between average model output and NASA TLX. In both cases, the summary statistics were used to assess how closely mean classifier output for each trial resembled its respective NASA TLX rating. A secondary objective was to assess the degree to which model accuracy changed as contextual data were included as model inputs alongside the physiological data inputs. A graphical plot is provided for each of the two model evaluation metrics along with discussion of observable trends. Each figure includes results on a per-participant basis across the two modeling approaches and both with/without context, for a total of four statistics per participant.

3 Results and Discussion

The two model evaluation metrics used in this analysis were: (1) correlation between average model output and NASA TLX; and (2) absolute difference between average model output and NASA TLX. For the correlation analysis, we believed it would be most suitable to derive a Pearson's correlation coefficient (r) on a per-person basis to better reflect how a given model tends to track any given person's cognitive workload across each trial. As such, the correlation coefficients for each of the six individual test participants and for all four workload classifications are illustrated in Fig. 2.

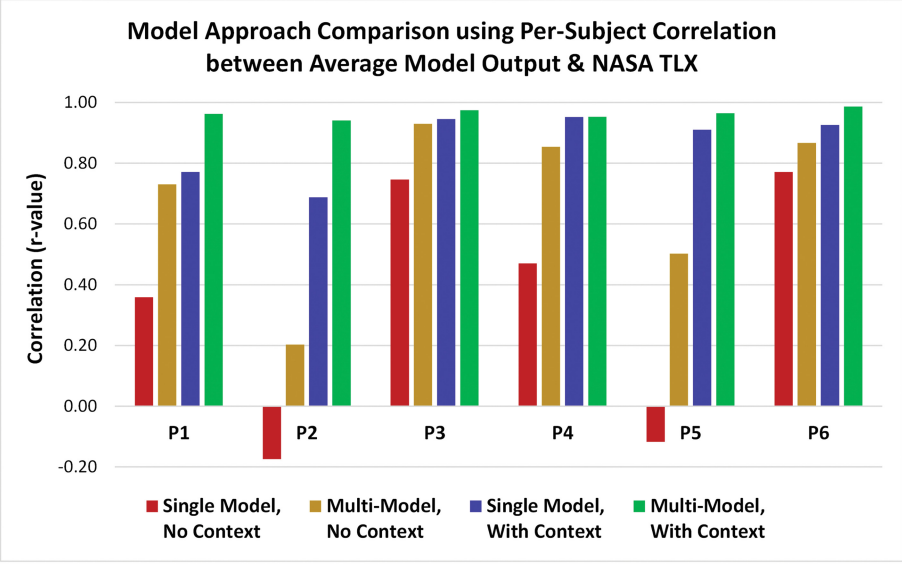

Fig. 2. Comparison of workload classification accuracy for six test participants using correlation between NASA TLX and average model output. (Color figure online)

The model approach with the highest correlations overall was the multi-model, with-context approach, which produced notably high correlations with $r = 0.94$ or higher for all six test participants. Perhaps the most noticeable finding in Fig. 2 is the consistent trend of improvement exhibited when shifting from a single-model approach to the multi-model approach, and similarly, from a no-context model to a with-context model. This trend occurs for all six test participants, albeit with varying degrees of improvement. This improvement is especially promising given the low correlations observed for participants P2 and P5 when using a single-model, no-context approach. The improvements in correlation were expected as a function of adding contextual data inputs. The effect of using a multi-model approach was less certain, though shifted in the hypothesized direction. From what can be observed in this small sample size, the magnitude of improvement for the multi-model approach was greater with the no-context models, as only minor improvements occurred for the with-context models. This may be due to the fact that with-context models already produced high correlations with a single-model approach. This finding may imply that for work environments in which contextual data inputs to workload classifiers are feasible, a multi-model approach might not be needed. However, if contextual data is not feasible as a model input, using a multi-model approach could produce substantial benefits.

The second model evaluation metric is the absolute difference between average model output and NASA TLX for each trial. This metric provides insight into each model's ability to produce workload classifications that accurately reflect the overall workload induced over the course of an entire trial. In contrast to Fig. 2, lower values shown in Fig. 3 indicate a greater degree of model accuracy by having a smaller difference from the desired NASA TLX value.

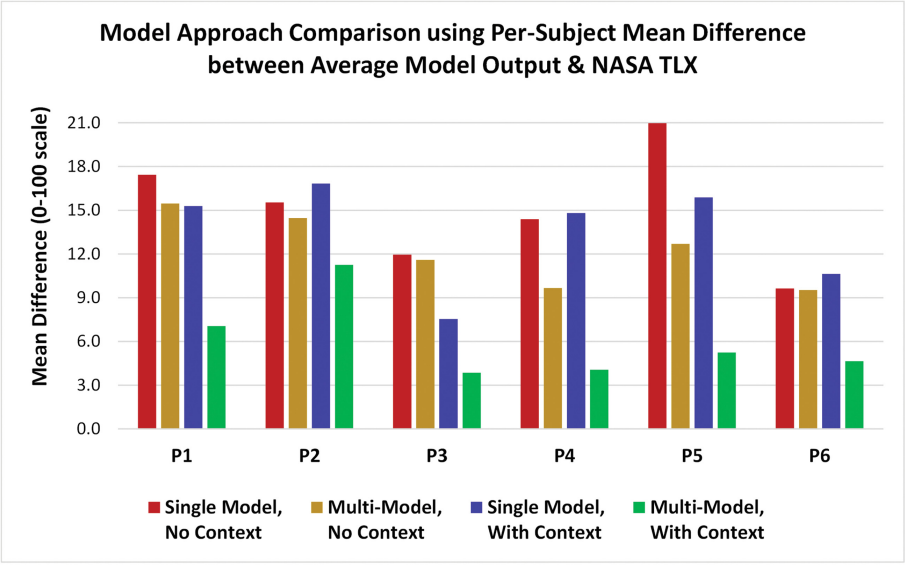

Fig. 3. Comparison of workload classification accuracy for six test participants using mean difference between NASA TLX and average model output. (Color figure online)

As shown in Fig. 3, the greatest accuracy (i.e., smallest difference from NASA TLX) was produced by the multi-model, with-context approach. On the 0-100 NASA TLX, this modeling approach resulted in a mean difference of less than 5.0 for four of the six test participants, with the other two participants having mean differences of 7.1 and 11.2 respectively. While this finding is generally consistent with the correlation statistics shown in Fig. 2, there are several additional key observations found in Fig. 3. Most notably, the magnitude of improvement using the multi-model approach is noticeably higher compared to the improvement provided by adding contextual data inputs. In five of the six test participants, shifting from a single-model, with-context approach to a multi-model, with-context approach reduced the mean difference by 50 % or greater. Further asserting the value provided with a multi-model approach is that a multi-model, no-context approach met or exceeded the single-model, with-context approach in five of the six test participants (P3 being the only exception).

4 Conclusions

In summary, we conclude that the utilization of a multi-model approach within our UAS task environment generally enhanced the FuSE^2 system's ability to accurately classify workload for 90 new trials across six test participants. This trend is observed for all six test participants across both model evaluation metrics (correlation & mean difference between NASA TLX and average model output) and occurs regardless of whether contextual data are included as model inputs. These results support our theory that relying on a single model classifier to produce consistently reliable estimates of human functional state presents risks, and that risk can effectively be mitigated through a diversified multi-model approach that is robust against the failure of any single model. We hypothesize the underlying cause of this potential improvement is due to the multitude of possible ways to process any given data set, and as such, each approach carries benefits and drawbacks that can never fully capture the entire picture in isolation. By blending a variety of approaches together, the complete picture can be more fully interpreted from multiple different angles and perspectives.

It is important to emphasize several key points to the multi-model approach that we believe influenced the promising results shown here. First, the human state assessments produced by the multi-model approach were driven bottom-up by the underlying NASA TLX sub-scale models. One potential drawback to the single-model approach is that workload, as the NASA TLX defines it, can be driven by different factors at different times, which may account for fewer discernible patterns to be discovered when focusing on the final aggregated NASA TLX value. We hypothesize that the FuSE^2 model classifiers were able to discover a more consistent pattern in the physiological and contextual training data sets by exposing the models to a more specific, low-level construct, as found in the NASA TLX sub-scales (namely, mental demand, physical demand, effort, frustration, temporal demand, and performance). Hence, it is possible that simply training a library of different models to the final aggregated NASA TLX value and fusing these results may not produce greater accuracy than a single-model approach. Another key point of emphasis is that each sub-scale model was trained using a different source

of second-by-second noise that is theoretically most appropriate for each respective sub-scale construct. The mental demand sub-scale model, for example, used EEG data to drive this noise, whereas the physical demand sub-scale model used ECG data, as dictated by empirical research on these constructs.

Although these results are promising, future research is needed to more thoroughly investigate the multi-model approach with other human functional states and within different operator task environments. Additional research is also needed to investigate and compare the difference in model accuracy between the bottom-up developments of a human functional state classification approach (as was done in this analysis) versus training and fusing multiple models to assess the same construct. Next, further analysis must be done to assess the multi-model classifier approach on a second-by-second basis, rather than solely the aggregated classifier results across entire trials. Lastly, the present analysis and follow-on research needs should be performed with other combinations of physiological data features – in particular, non-EEG models that may also include blinks, saccades, and facial expressions, to name a few – that may be more appropriate for specific work environments of interest.

Acknowledgement. Distribution A: Approved for public release. 88ABW Cleared 01/25/2016; 88ABW-2016-0243. This material is based on work supported by AFRL under Contract FA8650-11-C-6236. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL.

References

1. Blackhurst, J., Gresham, J., Stone, M.: The quantified warrior: how DoD should lead human performance augmentation. Armed Forces J. **4**, 11 (2012)
2. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors **45**(4), 635–644 (2003)
3. Wang, Z., Hope, R.M., Wang, Z., Ji, Q., Gray, W.: Cross-subject workload classification with a hierarchical bayes model. NeuroImage **59**(1), 64–69 (2012)
4. Gevins, A., Smith, M.E., Leong, H., McEvoy, L., Whitfield, S., Du, R., Rush, G.: Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. Hum. Factors **40**, 79–91 (1998)
5. Hoover, A., Singh, A., Fishel-Brown, S., Muth, E.: Real-time detection of workload changes using heart rate variability. Biomed. Sig. Process. Control **7**, 333–341 (2012)
6. Just, M.A., Carpenter, P.A.: The intensity dimension of thought: Pupillometric indices of sentence processing. Can. J. Exp. Psychol. **47**(2), 310–339 (1993)
7. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G.: Discriminating stress from cognitive load using a wearable EDA device. Technology **14**(2), 410–417 (2010)
8. Durkee, K., Geyer, A., Pappada, S., Ortiz, A., Galster, S.: Real-time workload assessment as a foundation for human performance augmentation. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2013. LNCS, vol. 8027, pp. 279–288. Springer, Heidelberg (2013)
9. Durkee, K., Pappada, S., Ortiz, A., Feeney, J., Galster, S.: Using context to optimize a functional state estimation engine in unmanned aircraft system operations. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS, vol. 9183, pp. 24–35. Springer, Heidelberg (2015)

10. Narendra, K.S., Balakrishnan, J.: Adaptive control using multiple models. IEEE Trans. Autom. Control **42**(2), 171–187 (1997)
11. Rowe, A.J., Liggett, K.K., Davis, J.E.: Vigilant spirit control station: a research testbed for multi-UAS supervisory control interfaces. In: Proceedings of the 15th International Symposium on Aviation Psychology, Dayton, OH (2009)
12. Finomore, V., Popik, D., Dallman, R., Stewart, J., Satterfield, K., Castle, C.: Demonstration of a network-centric communication management suite: multi-modal communication. In: Proceedings of the 55th Human Factors and Ergonomics Society Annual Meeting, HFES, Las Vegas (2011)
13. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Peter, A.H., Najmedin, M. (eds.) Advances in Psychology, vol. 52, pp. 139–183. North-Holland, Amsterdam (2006)