AFRL

TOWARDS A BETTER SCORING

IVAN J. TASHEV*, R. MICHAEL WINTERS *, YU-TE WANG *,
DAVID JOHNSTON *, JUSTIN ESTEPP†, NATHANIEL BRIDGES†

* MICROSOFT RESEARCH LAB – REDMOND, WA
† AIR FORCE RESEARCH LABORATORY – AFRL, DAYTON, OH

Disclaimer

I am a contractor working with Air Force Research Lab 711th Human Performance Wing and the views expressed in this presentation are my own and do not necessarily reflect the views of the Air Force or Department of Defense

Assumptions about the training process

- The goal is to improve the skill of the trainee to perform given task
- There are scenarios with various difficulty for the same task
- Training process consists of small indivisible trials
- In each trial is performed one scenario with given difficulty
- After each trial is computed a performance score (subject of this paper)

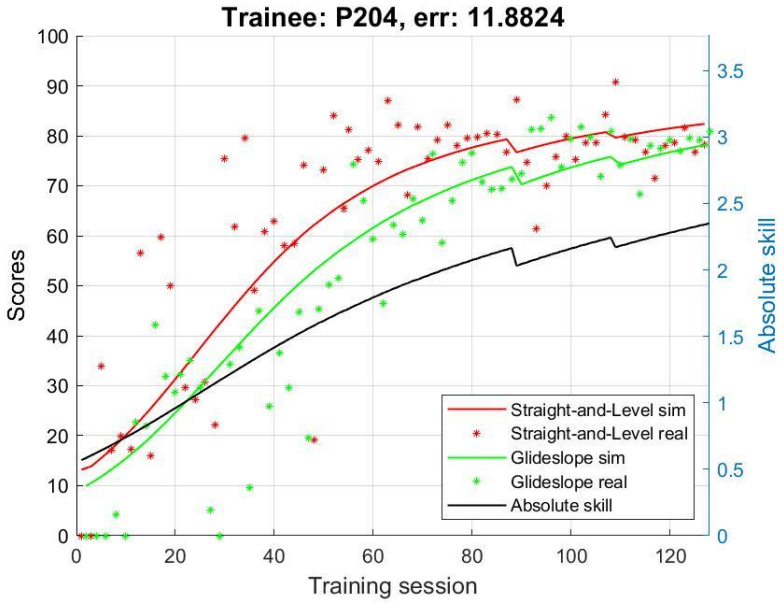
Tasks and current scoring

- Environment: flight simulator training
  - Three screen mode, or
  - Virtual Reality mode
- Task: straight line flight
  - Straight-and-level – maintaining constant course, speed and altitude
  - Glideslope – maintain constant speed and course approaching the runway
  - Duration is 2-3 minutes, variations in visibility, wind, thermals
- Current scoring
  - Flight simulator logs based
  - Averaged RMSE error from the prescribed straight-line flight and speed, scaled 0-100
  - Can be generalized to weighted sum of the normalized parameters: $S = 100\left(\sum_{i=1}^{N} w_i\left(1-\sigma_{i_{norm}}\right)+w_0\right)$
- Problems to address
  - Task dependent scoring!
  - Large number of non-informative negative scores with inexperienced trainees

Image owner: Microsoft via a contract with the Air Force

Proposed addressing of the issues

- Task dependency
  - Add additional parameters from the flight logs that are less task dependent: normalized deviation of the throttle and stick movements
- Treat the problems as a machine learning problem
  - Use the simulated scores as labels
  - Correlation with the parameters in the table
- Proposed classifiers
  - Linear regression
  - Support Vector Machines (SVM), in regression mode
  - Deep Neural Network (DNN), in regression mode
  - Extreme Learning Machine (ELM) in regression mode



Parameter	Corr. coef.
Airspeed_RMSENorm	-0.6751
PlaneAltitude_RMSENorm	-0.7413
LOCNeedle_RMSENorm	-0.6105
GSNeedle_RMSENorm	-0.3334
ThrottlePosition_STD	-0.2255
YokeXIndicator_STD	-0.2293
YokeYIndicator_STD	-0.2578

Ivan Tashev, R. Michael Winters, Yu-Te Wang, David Johnston, Alexander Reyes, Justin Estepp. "Modelling the Training Process", IEEE RAPiD 2022, September 2022

Image owner: Microsoft via a contract with the Air Force

Dataset, Training, and Results

- Dataset: 34 subjects, 11 scenarios, 1290 sessions
- Features:
  - The original four features
  - These above + the three control variations
  - Controls variations only
- Training:
  - Seven subjects with 90+ scores
  - One subject for testing, one for validation, the rest for training
- The results are average of all possible 42 combinations
  - Numbers are RMSE, lower is better

RMSE of the proposed approaches

Algorithm	Validation	Test
Baseline	0.5128	0.5128
Linear	0.1668	0.1952
SVM	0.1942	0.2052
ELM	0.1030	**0.1145**
DNN	0.1890	0.1960

Image owner: Microsoft via a contract with the Air Force

RMSE of DNN and ELM with various features

Feature set	Valid. DNN	Test DNN	Valid. ELM	Test ELM
Original	0.1928	0.1856	0.1151	**0.1159**
Orig.+contr.	0.1619	**0.2002**	0.3200	0.5301
Controls	0.2328	**0.2694**	0.5322	0.3566

More Results

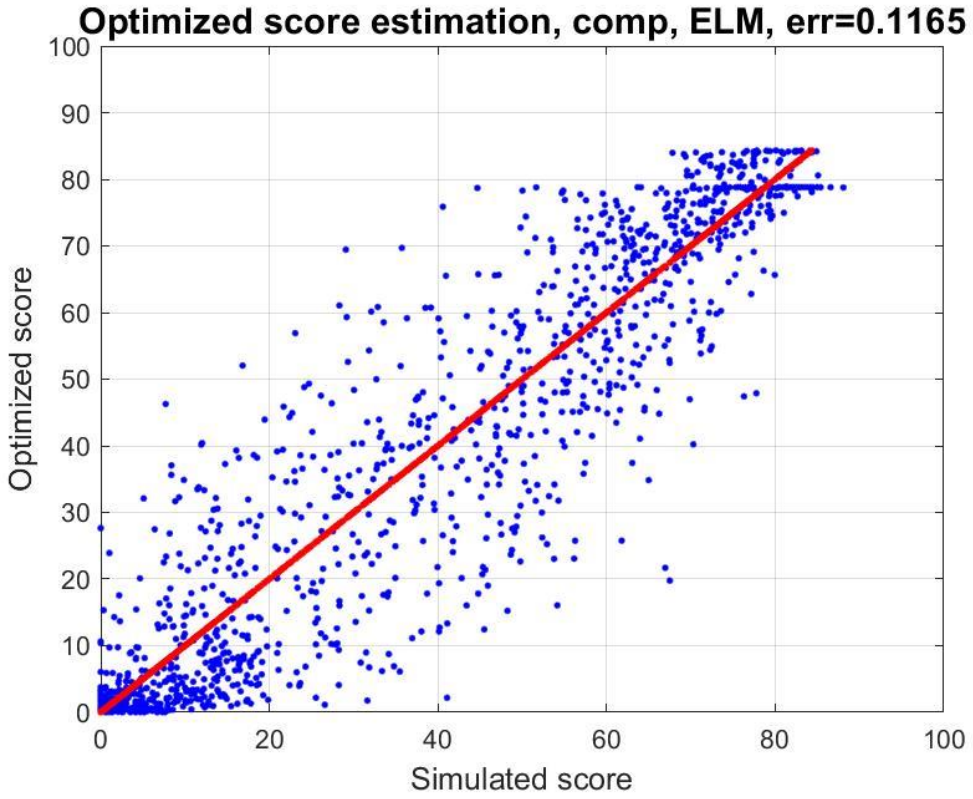


Image owner: Microsoft via a contract with the Air Force

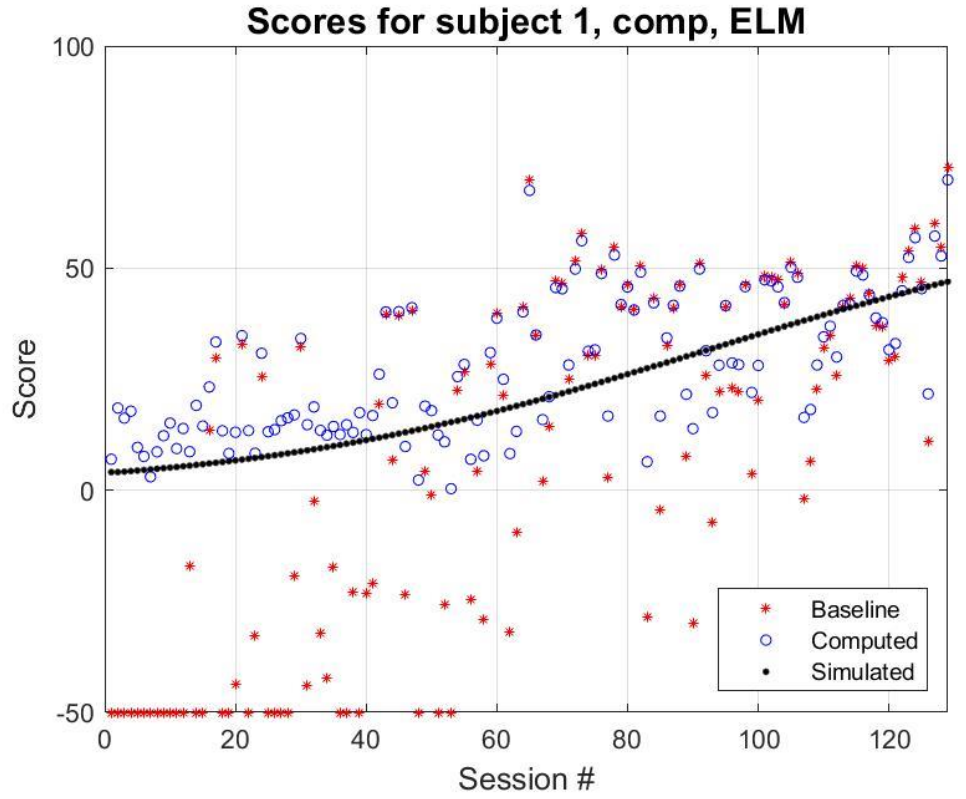


Image owner: Microsoft via a contract with the Air Force

Conclusions and Future Work

- Conclusions
  - The new ML-based scoring is better and more consistent
  - The new task independent features did not bring much to the table
  - ELM provides the best results on the original feature set, DNN seems more robust on all three
  - The labels are good reflection of the subject's cognitive load
- Future work
  - Try the same approach with physiological data (EEG, gaze, ECG, breathing, etc.)
  - The goal is to make the scoring person- and setup- independent

QUESTIONS?