A Model-Based Measure to Assess Operator Adherence to Procedures

Alexander J. Stimpson[1]
Luisa S. Buinhas[2]
Scott Bezek[1]
Yves Boussemart[1]
M.L. Cummings[1]
[1]Massachusetts Institute of Technology, Humans and Automation Lab
[2]Delft University of Technology

Procedures play an important role in domains where humans interact with critical, complex systems. In such environments, the operator's ability to correctly follow a given set of procedures can directly impact system safety. A quantitative measure of procedural adherence during training for complex system operation would be useful to assess trainee performance and evaluate a training program. This paper presents a novel model-based objective metric for quantifying procedural adherence in training. This metric is sensitive to both the number and nature of procedural deviations, and can be used with cluster analysis to classify trainee performance based on adherence. The metric was tested on an experimental data set gathered from volunteers using aircraft maintenance computer-based training (CBT). The properties of the metric are discussed, along with future possibilities.

INTRODUCTION

Procedures are commonplace for guiding interactions between humans and complex systems. Standard operating procedures (SOPs) are typically developed to improve system safety and operations by providing operators a specific set of actions appropriate for various circumstances. To prepare for procedure-based environments, trainees often complete a rigorous training program that includes practice utilizing the procedures. Typically, the ability of trainees to follow SOPs is determined qualitatively by a training supervisor, who often only considers the final outcome of each training module (such as in surgical training, (Lammers et al., 2008)). A quantitative measure of procedure adherence for trainees undergoing instruction for a complex system would be useful to assess trainee performance as well as evaluate a training program.

This paper presents a novel model-based objective metric for quantifying procedural adherence in training. By generating a score value for each trainee, the Procedure Adherence Metric (PAM) can augment the assessment of a training supervisor with an objective performance evaluation. This metric is sensitive to both the number and nature of procedural deviations, and can be used with cluster analysis to classify trainee performance based on adherence.

The metric was tested on an experimental data set gathered from volunteers using an aircraft maintenance computer-based training (CBT) interface. Based on this data set, the main properties of the PAM score were identified, and the metric was used to group trainees using clustering algorithms. In this paper, the development and properties of the metric are discussed, along with the current applications and further possible uses of the metric to predict future performance of trainees.

LITERATURE OVERVIEW

A large body of research has attempted to answer the question of how to assess the performance of a trainee. Assessment strategies typically depend on the type of knowledge to be obtained in training, which can be split into procedural (how to perform a task) or declarative (improved understanding of a topic) strategies (Alessi, 2000). Within the scope of this paper, the emphasis is on the assessment of training for procedural knowledge.

Most training performance assessment techniques fall into two categories: qualitative assessment and performance-based assessment (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007). Qualitative assessment usually involves evaluation by an expert instructor either in training (Govaerts, et al., 2007) or after a training session (Owen, Mugford, Follows, & Plummer, 2006). However, qualitative assessment methods have been challenged for their lack of accuracy and reliability (Govaerts, et al., 2007).

Performance-based assessment adopts a more objective approach, where trainees are expected to operate under a certain set of performance criteria (Hamman, 2004). Performance-based methods typically utilize simulation or examination to directly test these aspects of performance. In simulation, embedded assessment tools can provide an unobtrusive way to collect data that correspond to the learning objectives of the training (Nählinder, Oskarsson, Lindahl, Hedström, & Berggren, 2009). Since in- and post-training performance are not always directly correlated (Ghodsian, 1997), both qualitative and performance-based techniques can have difficulty in evaluating the operational applicability of the training (Bjork & Bjork, 2006).

Researchers have made clear that further research is needed to identify new models and methods for assessing procedural skills in simulation (Lammers, et al., 2008). A new quantitative method based on how well trainees follow procedures during training could help in overcoming these challenges and support superior predictions of future performance, both in training and the real-world.

Most previous research on procedure adherence only accounts for the number of procedural deviations (e.g. (Gerbaud, 2008)). However, not all procedural deviations necessarily degrade performance. For example, experts sometimes reorganize or skip steps in commonly-used procedures to solve

problems more efficiently. Since deviations can arise for several reasons, the nature of a procedural deviation must be taken into account in any assessment. Lammers *et al.* (2008) propose that the characteristics and the components of each procedure must be clearly defined before constructing an evaluation tool. Vague or unclear terms may lead to non-compliance (Rogovin, 1979), which may lead to the assumption that procedural deviation relates to trainee inability to perform a task.

Given these considerations, an inexpensive, embedded performance-based metric for procedural adherence is needed. Ideally, we propose this metric would have the following characteristics:

1. Sensitive to number and type of deviations from the prescribed procedure
2. Able to compare trainees on procedural adherence
3. Able to classify trainees into subgroups (i.e., distinguish good from poor trainees)
4. Simple to interpret
5. Predictive of operational performance

The following section discusses a variety of possible techniques that could be applied to measuring procedural deviations, and compares how these techniques fit the ideal characteristics of a performance-based metric.

METHODS

Procedure as a Sequence

An SOP defines a series of actions for the user to take, typically under a certain set of initial conditions that make the procedure appropriate. The set of actions contained in a procedure can be translated into a sequence, with each action having a previous action and a subsequent action (Figure 1). Following a procedure similarly consists of an ordered sequence of actions, generated as a trainee attempts to complete the procedure. However, the sequence of trainee actions may not exactly match the sequence given by the SOP. Errors such as omission of an action, performing actions out of order, or substitution of an action with an incorrect one can create mismatches between the procedure sequence and the trainee sequence. Typically in a complex system, there are more actions available to the user than are needed for any particular procedure. Therefore, it is even possible for a trainee to generate actions that are never observed in the procedure sequence.

In this framework, procedure adherence can be measured by the difference between a SOP sequence and the trainee sequence. Numerous methods that measure the distance between sequences have been developed, including sequence-based methods (e.g. Levenshtein distance (Levenshtein, 1966)) and model-based methods (e.g. Kullback-Leibler divergence (García-García, Hernández, & Diaz de Maria, 2009)). Sequence-based methods focus on the direct comparison of the sequences, while model-based methods model each sequence and then compare the similarity of the models as a proxy for sequence distance. To select the best method for the calculation of sequence distance, four elements for procedure adherence measurement were considered:

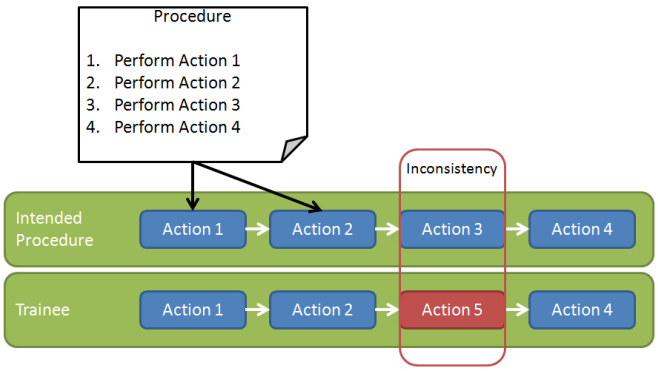

Figure 1. An example procedure sequence and a corresponding trainee-generated sequence

1. Ability for different types of errors to be penalized separately (i.e. error weighting)
2. Non-equal sequence length between the observed and expected sequence of actions, i.e., where the user performs a different number of actions than the prescribed procedure
3. Sensitivity to number of deviations between the observed versus expected sequence of actions
4. Sensitivity to trainee action order – the ability to distinguish between correct and incorrect ordering

Three methods that calculate sequence distance were considered: Two sequence-based methods (Levenshtein distance and suffix arrays (Manber & Myers, 1990)) and one model-based method (Kullback-Leibler divergence). We chose the Kullback-Leibler (KL) approach as it meets all four criteria above. The KL divergence inherently creates a variable penalty for user deviations from the intended procedure based on transition probabilities between actions. This allows the metric to be sensitive to the order of the actions in the sequence. Because of this advantage, the KL divergence was used as the basis for the procedural adherence metric. However, it required adaptation for our purposes, described below.

Proposed Metric

The main goal of measuring procedure adherence is to assess trainees' performance against the SOP. Additionally, trainees can be objectively compared against each other based on their training performance, and tracking procedure adherence can indicate struggling trainees that need re-training.

Our proposed metric, the Procedure Adherence Metric (PAM) was based on the KL divergence between sequences. A single final value can be calculated for an entire sequence, or the deviation can be calculated at each action in the trainee sequence (Figure 2). To identify whether a single KL divergence value or the progression of the divergence over time is a better measure of overall procedure adherence, several features of the KL divergence over the course of a typical training module were analyzed. These included final value, mean divergence, area under the curve, peak divergence, peak-final difference, and peak-final ratio. The divergence at each action was calculated by comparing the model of the subsequence,

represented by all events up to that action to the intended procedure. It was determined that using the area under the KL curve was the most useful form of the metric because it provided the closest rankings of trainees (from best to worst) as from an expert evaluator. Using the area indicates that a score value φ can be computed as:

$$\varphi = \sum_{i=1}^{N} D_{KL} \qquad (1)$$

Where N is the number of events (interactions) in the training sequence, and D_{KL} represents the symmetrized Kullback-Leibler divergence between the trainee sequence of states 1…i and the intended sequence of states of equivalent length. If N is greater than the number of states in the intended sequence (M), the complete intended sequence is used for all i>M. It is important to recognize that as the PAM is based on divergence, a lower score indicates better performance.

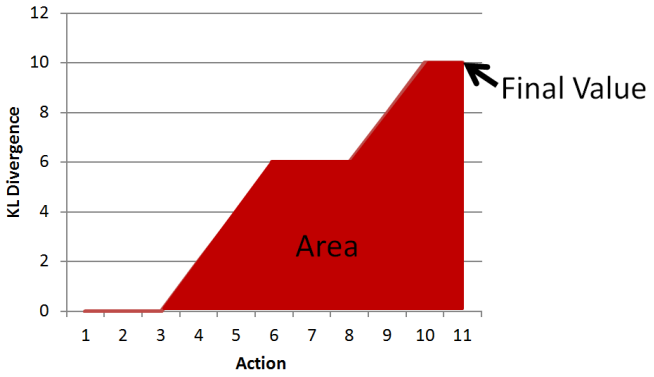

Figure 2. Illustration of aspects of KL divergence progression that could be used in the measurement of procedure adherence

A potential issue that arises in the use of the KL divergence for the PAM is zero-probability values in the transition matrix. This transition matrix represents the probability of all transitions between states in the model. For a model based on a sequence, the maximum likelihood estimate simply counts the number of times a consecutive state pair is found (e.g. state 1 to state 2), and normalizes by the number of transitions. If a particular set of states are never observed consecutively, the count (and therefore the probability estimate) for that transition is zero. The size of the state transition matrix is heavily dependent on the number of existing states (NxN for N states), and can be large for CBT settings.

Often the set of actual transitions in any particular training procedure will not cover the entire set of possible transitions. When included in the model, these zero probability events send the KL divergence to infinity. Instead a small (but non-zero) probability can be assigned to transitions that do not occur in the intended procedure. This results in a large divergence score (poor performance) in the PAM, but does not send the divergence to infinity. Smoothing methods can assign these low probabilities in the transition matrix. Smoothing (or frequency estimation) techniques can be employed to estimate the probability of occurrence of novel events, such as in a Markov Chain. For the PAM, Good-Turing smoothing was selected as the smoothing method based on its ability to handle large numbers of novel events and applicability to first-order Markov models.

Application of PAM to CBT

In a CBT setting, the trainee's interactions with the interface (mouse clicks, typing, etc.) represent an important objective data source that can be used to quantify performance. In these settings where a large number of actions are available, sparse transition matrices for both the procedure and trainee sequences can arise. In general, the problem of sparse transition matrices can be alleviated by grouping several events into a single state to reduce the total size of the transition matrix. In the context of procedural adherence in training, functional training objectives for a module can be used to create the groupings. All events that correspond to a single training objective can be consolidated into a single state. For example, all sub-tasks related to the objective of turning off a machine could be grouped under a single "Shut-down" state. This would reduce the entire set of states represented by each individual action into a single state indicating that the trainee is pursuing the correct objective.

The sequence generated by a user's behavior with a CBT can then be compared to the prescribed sequence via the PAM. To test the usefulness of the metric in evaluating performance and procedural adherence in CBT settings, data was collected on a sample CBT interface. The process is detailed below, including an example of state consolidation and the results of the metric application.

DATA COLLECTION

The experimental population comprised 17 volunteers, 12 males and 5 female. The participants' ages were between 18 and 57 years old (mean 27.3 years, standard deviation of 9.3 years). In the data collection exercise, volunteers were asked to resolve a particular aircraft maintenance-related problem. The computer-based interface used was created by Boeing Flight Services with the intent of training maintenance personnel to resolve problems with aircraft systems.

The volunteers were asked to solve a simulated maintenance task using the CBT. Their role was to investigate an error message provided at the beginning of the session. The volunteers were asked to perform an identical task twice using the CBT, with no time pressure. Subjects were instructed to solve an aircraft error message by following a series of diagnostic steps, including error identification, gathering reference information, solution identification, and solution implementation. A walkthrough of the interface was given before the first run. The first run acted as a familiarization with the CBT interface, and the second run was used for data analysis. A log file was generated for each run that contained a list of user interactions with the interface. To reduce the size of the transition matrix, groupings of actions were treated as single states based on cognitive similarities between the actions. These groupings were 1) Main Menu, 2) Searching for Message, 3) Performing Tests, 4) Engine Indication and Crew Alerting

System Message, 5) Inaccessible Menus (error), 6) Fault Code Index, 7) Task Manual, 8) Performing Repairs, and 9) Corrective Action.

RESULTS & DISCUSSION

The difference between the sequences of events generated by the participants and the intended procedural sequence was measured using the PAM. For the second run (data analysis run), the KL divergence was computed at each state transition. Figure 3 shows the divergence graph for all participants by interaction across the experimental run, and Figure 4 shows the overall PAM score for each participant. The divergence is plotted by transition count, with each value representing the divergence at the user subsequence of that number of transitions. For example, at the time the user has performed 20 actions, there are 19 transitions that will be included in the divergence calculation. Time to task completion is not considered in this analysis. Three outlying performances are seen, representing trainees that became lost in the interface.

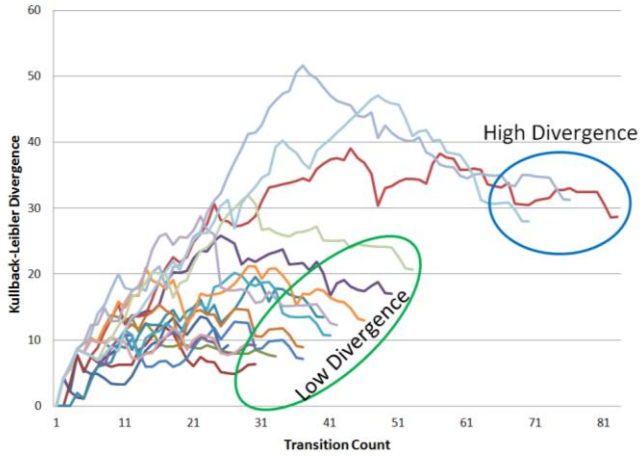

Figure 3. KL divergence over the course of the training run

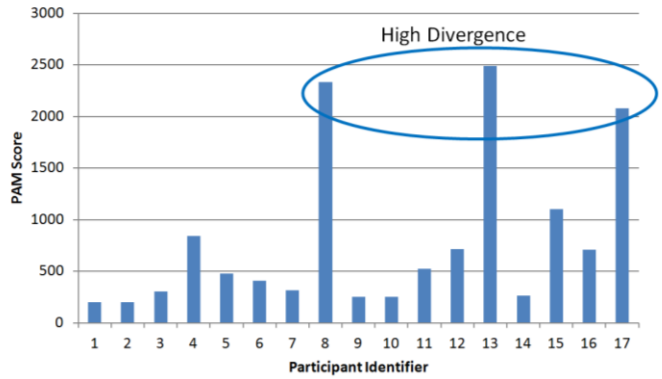

Figure 4. PAM score based on area under the curve for each participant

The divergence score can take on a value between 0 and infinity. A score of 0 represents no difference between the user and intended model. The upper bound of the PAM value is based on 1) the smoothing method and 2) the number of transitions in the trainee sequence. The meaning of any particular value of the PAM score is dependent on both the training task and objectives. In some systems, it may be imperative that every step of a procedure is followed to the letter. In these cases even modest deviation scores may indicate a need for retraining. In other circumstances, only the poor performers may need remedial work. Thus the interpretation of both the numerical score and the significance of the deviations are dependent on the specific procedure under evaluations.

At each transition in Figure 3, the increase, constant, or decrease of the divergence value indicates something about the behavior of the user. Two types of divergence can be identified and separated: performing the right action at the wrong time, and performing a wrong action. Further investigation showed that an increase in the KL divergence at a transition indicated a deviation from the SOP at that time, while a decrease indicated that the user had performed a task too early. A level portion of the graph indicates that the user is following the intended procedures. Since the PAM incorporates the area under this graph, it can differentiate between an incorrect action and an action performed out of order. However, there are cases where performing a correct action at an incorrect time can be just as dangerous as a totally incorrect action. Work is underway to determine how to add weighting parameters that include heavier penalties for performing critical actions incorrectly.

In order to investigate how the metric addresses the timing of an error, two artificial sets of sequences were created. One contained a set of errors early in a procedure, and the other contained an equivalent set of errors later in the procedure. The results indicated that the divergence score applies a higher penalty to an early deviation than a later one. Thus, trainees' mistakes earlier are penalized more than later mistakes. We are currently investigating potential weighting factors that could modify the balance of the penalties for early mistakes compared to later mistakes.

An additional use for the metric is grouping users of similar behavior in order to identify strong and weak performers. In training settings, it may be efficient for the supervisor to be able to identify a subset of trainees that all exhibit similar procedural adherence deficits, such that these trainees could undergo the same retraining process, or to highlight difficult parts of a procedure. These potential groupings could be determined by cluster analysis. A simple cluster analysis for the data in Figures 3 and 4 revealed two clusters that separated the outliers from the other trainees. While this may seem apparent by inspection for this data set, the ability to objectively determine groupings could be useful in providing further feedback to a supervisor. The cluster analysis used here was based entirely on the divergence data, but the consideration of demographic inputs such as computer experience or CBT familiarity could help inform the clustering process.

With more data, additional groupings based on behavioral similarity could be revealed using this clustering methodology. These outliers showed behaviors that were indicative of being 'lost' in the interface, such as consistently clicking on menus or buttons that were unrelated to the task. Despite receiving training on the exact procedure to be followed, these subjects had difficulty in regaining the correct solution path once they became lost. All subjects were allowed to complete

the task, and the behavior of the outlying subjects seems consistent with a trial and error strategy.

There are several practical applications of the PAM. The PAM makes use of not just the outcome of each training module, but all of the training process data. This approach objectively allows a supervisor to identify possible stumbling points throughout each training module, either as a result of poor training design or problems with the individual trainee. By providing an objective performance score for trainees based on their adherence to the SOP, supervisors can augment their own view of the trainees' strengths and weaknesses. While not intended to replace the supervisor's assessment, the PAM and the underlying KL graph give a supervisor insight into which sections of the procedure are difficult.

CONCLUSIONS AND FUTURE WORK

PAM, a new metric for measuring procedure adherence in CBT environments, provides an objective measure that can be used to compare trainees' procedural adherence. As has been demonstrated, it is sensitive to both the number and nature of deviations from intended procedures. Cluster analysis can classify trainees in groups based on their adherence. By combining procedural deviations into a single metric, PAM provides a simple method for supervisors to judge trainees' adherence.

Because the PAM is based on the Kullback-Liebler divergence method, there are some limitations. Of most concern is the inability to weight different error types, as well as weighting them based on when they occur in the procedure. These areas are currently under investigation.

Further research can be conducted that may improve the PAM. Additional larger data sets will help to validate its usefulness in classifying trainees. Training literature generally considers time to task completion as an indicator of training and operational performance, and the inclusion of temporal information may bolster the metric's ability to make accurate operational performance predictions.

Future work can also expand the applications of the PAM. The metric could be used to evaluate training modules based on the aggregate behavior of trainees. If a particular module resulted in poor PAM scores across all trainees, this may be indicative that the training module is poorly designed. Use of the metric to predict operational performance is also an important potential application of this research.

Acknowledgements

This research was conducted with support of Boeing Training and Flight Services. The authors thank the participants who volunteered to participate in the exercise.

References

Alessi, S. (2000). Simulation design for training and assessment *Aircrew training and assessment* (pp. 199-224). New Jersey: Lawrence Erlbaum Associates, Inc.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse *Memory and Society: Psychological Perspectives* (pp. 109-133): Psychology Press.

Blessing, S. B. A., J.R. (1996). How People Learn to Skip Steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 576-598.

García-García, D., Hernández, E. P., & Diaz de Maria, F. (2009). A new distance measure for model-based sequence clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(7), 1325-1331.

Gerbaud, S. M., N.; Ganier, F.; Arnaldi, B.; Tisseau, J. (2008). *GVT: a platform to create virtual environments for procedural training.* Paper presented at the IEEE Virtual Reality, Reno, Nevada.

Ghodsian, D. B., R.A.; Benjamin, A.S. (1997). Evaluating Training During Training: Obstacles and Opportunities. In M. A. Q. a. A. A. Ehrenstein (Ed.), *Training for 21st Century Technology: Applications of Psychological Research* (pp. 63-88). Washington D.C.: American Psychological Association.

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in health sciences education, 12*(2), 239-260.

Hamman, W. (2004). The complexity of team training: what we have learned from aviation and its applications to medicine. *Quality and Safety in Health Care, 13*(suppl 1), i72-i79.

Lammers, R. L., Davenport, M., Korley, F., Griswold Theodorson, S., Fitch, M. T., Narang, A. T. (2008). Teaching and assessing procedural skills using simulation: metrics and methodology. *Academic Emergency Medicine, 15*(11), 1079-1087.

Levenshtein, V. I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals.*

Manber, U., & Myers, G. (1990). *Suffix arrays: a new method for on-line string searches.* Paper presented at the Proc. of the first annual ACM symposium on Discrete algorithms.

Nählinder, S., Oskarsson, P. A., Lindahl, B., Hedström, J., & Berggren, P. (2009). *Effects of simulator training: motivating factors*: Information Systems, Swedish Defence Research Agency (FOI).

Owen, H., Mugford, B., Follows, V., & Plummer, J. L. (2006). Comparison of three simulation-based training methods for management of medical emergencies. *Resuscitation, 71*(2), 204-211.

Rogovin, M. (1979). Three Mile Island: a report to the commissioners and the public. Washington, DC: Nuclear Regulatory Commission.

Vora, J., Nair, S., Gramopadhye, A. K., Duchowski, A. T., Melloy, B. J., & Kanki, B. (2002). Using virtual reality technology for aircraft visual inspection training: presence and comparison studies. *Applied Ergonomics, 33*(6), 559-570.