Bootstrapping Assessments for Team Simulations: Transfer Learning Between First-Person-Shooter Game Maps

Benjamin D. Nye[1][0000−0002−5902−9196], Mark G. Core[1][0000−0002−0438−3868], Sai V.R. Chereddy, Vivian Young, and Daniel Auerbach[1]

University of Southern California, Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista CA 90094 {nye,core}@ict.usc.edu

Abstract. Assessing teams and providing feedback on scenario-based training typically requires human observers or scenario-specific metrics crafted by experts, due to the complexity of general-purpose automated tools to assess team performance. Machine learning can help infer team performance patterns, but labeled data for a specific training scenario is often sparse. To address this issue, the Semi-Supervised Learning for Assessing Team Simulations (SLATS) project investigated the feasibility of semi-supervised learning and transfer learning which leverages training data from related scenarios to classify performance on a target scenario with the same metrics but a different terrain context. To this approach, we analyzed performance of teams in the first-person shooter Team Fortress 2 (TF2). TF2 teams for the "Capture Point" mode were classified into archetypes based on the performance of the team and the performance of individual members of the team across the corpus: novice, weak link, team of experts, and expert team. To investigate the feasibility of transfer learning, we isolated matches from two of the most frequent maps/terrains. Results found that leveraging data from the source map always improved classification F1-scores compared to relying solely upon target (test) map training data. The greatest benefits were observed when target data was limited (0 to 42 target examples).While further research is required to explore the effectiveness of transfer learning across training scenarios that are more dissimilar (e.g., different simulations, rather than just different maps), these results offer a promising direction to help bootstrap team assessments on new training scenarios by leveraging data from earlier, comparable scenarios. However, efficiently calculating reusable metrics for model features based on low-level scenario events and logs remains a challenge that requires further research.

Keywords: Semi-supervised learning · Transfer learning · Team training

1 Introduction

Assessing teams and providing feedback on scenario-based training is traditionally ad-hoc, due to a lack of general-purpose automated tools to assess team

performance. In many cases, the gold standard remains live observers with a virtual control panel or physical scorecard to record performance outcomes. While standards such as xAPI have facilitated the development of general-purpose data analytics [1], the patterns that represent expert versus novice performance can vary substantially based on scenario difficulty or objectives. Machine learning can help infer these patterns, but labeled data for a specific training scenario is often sparse.

To address this issue, the Semi-Supervised Learning for Assessing Team Simulations (SLATS) project investigated the feasibility of transfer learning which leverages training data from related scenarios to classify performance on a target scenario (i.e., same metrics but different conditions). This work aligns to efforts such as the Army's Synthetic Training Environment, which should enable scenario-based team training such as battle drills to be conducted in simulated environments with individual and team actions logged using the xAPI standard [4]. The SLATS project was driven by three goals:

1. **Classify Team Performance**: Automate or semi-automate activities that an observer-trainer might need to perform during or after a training scenario, to enable greater opportunities for team training (e.g., reducing cost and expertise bottlenecks). While not all abilities of a human observer or trainer can be replicated in an automated system, sufficient data should exist in a simulation to identify common errors that should be flagged as areas for improvement.
2. **Diagnose Performance Issues**: Develop a set of key team metrics and data views that aggregate lower-level scenario-specific assessments into actionable and interpretable insights. The use-cases of strongest interest are to produce metrics for: individual feedback, team feedback, scenario adaptation (for a future simulation), and instructor review/assessment.
3. **Generalized Framework**: Develop a re-usable set of metrics and tools that can be applied to assess team training in a variety of scenarios, leveraging industry and standards for recording performance and learning events.

In this paper, we investigate the potential of transfer learning to help achieve these goals. Analyses on an existing large corpus of team game scenarios (Team Fortress 2) are presented. We also present context on the overall machine learning pipeline used by SLATS, with an emphasis on how effectively these models could facilitate diagnostic feedback and generalization to new types of simulations. The results presented indicate that transfer learning offers an effective way to improve assessment for scenarios in a similar training system, but that the ability to generalize models broadly remains a challenge.

2 Background

2.1 Assessment Methodologies for Team Training

While scenario-based assessments have been explored in many educational contexts, assessments to support both team and individual learning remain challenging and are traditionally scenario-specific and labor intensive. Given the difficulty

to develop such computer-based assessments, they are frequently not used even when the training itself delivered using computer-based training. For example, large organizations such as the Army still rely primarily on live observer-trainers to watch the exercise and manually determine feedback and after-action-review items (e.g., sustain vs. improve priorities), limiting training feedback to times where human experts are available. However, effective training requires many practice opportunities; it is not feasible for a large number of teams to practice toward expert performance when experts must facilitate each session.

Team training is substantially more complex than individual training, because learners may vary not just by skill level but also by the types of skills they are expected to know (e.g., specialization). Research on team assessment and performance has proposed role-based models for team behavior to address these issues [2, 3]. There may also be differences in team versus individual performance. Metrics development has looked at distinguishing between i) individual tasks, ii) team tasks (outcomes), and iii) teamwork (process), in projects such as the Surveillance Scenario Team Tutor [3] and Squad Overmatch [5].

Smith-Jentsch, Johnston, and Payne [9] break teamwork into four categories: information exchange (domain-relevant content of communications), communication delivery (e.g., clarity, brevity, using proper terminology and language), supporting behavior (back-up behavior to correct errors or fill gaps), and leadership (adapting priorities and guidance to changes in the situation). Integrating across these frameworks, an ideal-world team assessment would account for: a) propagation of errors (e.g., inability to complete a task due to a teammate's failure), b) external influences (e.g., good process/bad outcome), and c) back-up behavior (e.g., assigning credit for successful performance to the proper individual). The models should also distinguish between task work (e.g., performance) vs. teamwork (e.g., coordination). These behaviors imply that assessing teams meaningfully requires capturing both team and individual metrics, with some structure or data-derived inferences to determine how individual metrics relate to higher-level team processes and outcomes.

2.2 Learning: Semi-Supervised and Transfer Leaning

To address the cold start problem for scenario-based training data, we are using a semi-supervised approach to build a classifier to detect engagement archetypes. Given that labeled data is often hard to collect, semi-supervised methods leverage a small amount of labeled data to make better use of a larger set of unlabeled data [10]. In earlier work by our group, the SMART-E project (Service for Measurement and Adaptation to Real-Time Engagement) applied semi-supervised learning for generalized, automated assessment of engagement by individual learners. Research with SMART-E found that metrics were able to generalize across systems for engagement [7] and that semi-supervised learning offered advantages for classifying and interpreting engagement archetypes such as distracted learners versus those racing through the content [8]. As such, a goal for SLATS was to generalize this technique to assessing teams in scenarios.

However, while investigating semi-supervised techniques, we recognized that our semi-supervised approach was primarily helpful for an initial scenario where archetypes were not yet well understood. Later scenarios should be much faster to classify accurately if transfer learning can boost new scenario assessments based on patterns in earlier well-analyzed scenarios. However, the benefit of transfer learning depends on the similarity between the tasks [6]. Even for different maps or variations of scenarios with the same objectives, different team behaviors might be more successful overall.

3 Approach

The SLATS architecture is designed to process data in stages as shown in Figure 1, such that each subsequent stage only relies on the prior stage as a data source. Raw events and logs are first produced by a training scenario, which are either directly recorded as xAPI statements or processed through a log-file converter to generate xAPI records. A log-cleaner function then fixes these raw logs to produce a second canonical xAPI log for processing (meaning that the raw xAPI statements always exist for alternate cleaning or record checks). In the second stage, the raw xAPI logs are analyzed to produce two types of metrics: direct metrics and intermediate metrics. Direct metrics require xAPI log data to perform their calculations (e.g., number of deaths for a player that session), while intermediate metrics can be calculated only based on other metrics (no xAPI data needed). Metrics may be individual or team, with certain team metrics being more likely to be intermediate (i.e., derived from the individual players). Metrics may either be custom functions or they may be determined by a lightweight markup file which specifies certain functions and aggregations (e.g., average, min/max, etc.).

As shown in Stage 3, a team session vector can be specified, which specifies the set of metrics that will be available as features for classifying team performance. In the example analysis below, teams are classified only on team-level metrics for easier interpretation, but this is not a requirement. A session vector is calculated for each scenario sessions, both for labeled data (known archetypes) and unlabeled data. In a multi-team match, each team will have its own session.

Classification occurs during the final stage. Following the approach described for the SMART-E semi-supervised model [8], unlabeled sessions are clustered based on their feature vectors. An alignment algorithm calculates the global best-match between each cluster and the labeled data for each archetype. Then, data for each cluster is assigned a candidate label based on the archetype which aligned to it. This pooled data set includes both truly labeled data and cluster-aligned data, which are used to train a machine learning model. Different clustering algorithms and classifier types may be selected using parameters, with Gaussian Mixture Models (GMM) clustering and Logistic Regression classification used by default. This approach to pooling labeled and unlabeled data for the classifiers increases accuracy versus using only unlabeled data for training and

exploratory analysis indicates benefits up to about four times as much unlabeled data as labeled data (e.g., 20 labeled vs. 80 unlabeled).

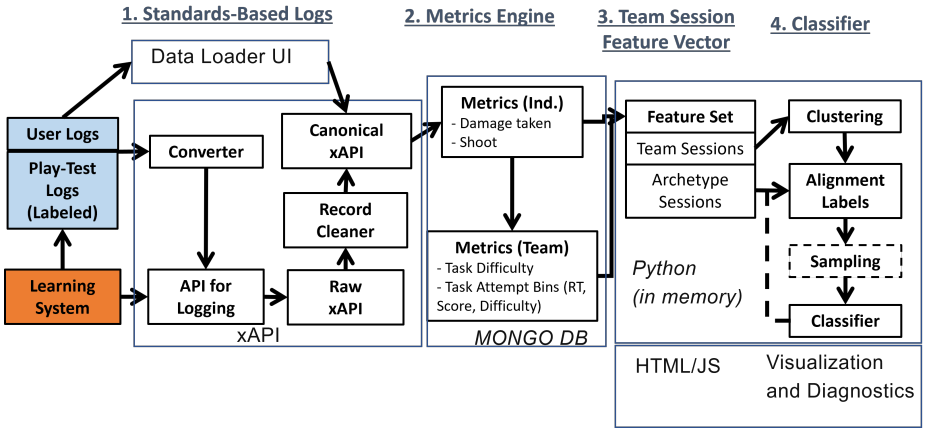

Fig. 1. SLATS Architecture Diagram

3.1 Team Archetypes

SLATS classifies teams into "archetypes" of performance that represent their performance level and use such classifications to areas to practice next. Unlike a traditional 0 to 100 score, we are instead interested in the development stage of a team from a poorly-coordinated set of novices to a highly-effective expert team. This is important for the ability to re-use metrics across different scenarios and simulations. For example, in one scenario it might be reasonable for an expert team to have only 10 communicative actions, while in another, an expert team might require 100. Moreover, metrics are not necessarily linear between archetypes: an expert gaming team might have fewer kills, because they win decisively without an extended conflict. Manually filling in and updating scenario-specific parameters and weights would be time consuming for scenario authors. To avoid this requirement, models were leveraged to estimate and update parameters with the goal of being able to distinguish between different classes of team behavior. Different archetype categories may be specified per-system that is registered in the SLATS framework. In the current work, we focused on classifying:

1. Expert Team: Team is effective and composed of successful individuals.
2. Team of Experts: Team members are good at individual tasks, but the team is not successful, such as due to poor communication or coordination.
3. Weak Link: Team members are good at individual tasks effectively, but the team is not successful, such as due to poor communication or coordination.

4. Novice: Team performances is poor, which is also reflected by lack of success or experience of its individual members.

Typically, a ground truth data set would be established based on expert labeling of a small set of sessions of each category. However, in this case due to the very large corpus of Team Fortress 2 (TF2) matches, we inferred labels based on knowledge about team performance and the individual performance of each team member across all their known matches. Each team (unique combination of individuals) was characterized by its team performance and its predicted performance based on a linear regression of team members' statistics across all their known matches (shooting, support, and survival).

Gold labels for teams were defined by the following heuristics, for the designated archetypes. While the broader SLATS project explored other archetype categories, research on transfer learning focused on these categories.

1. Expert Team: Over 75^{th} percentile team performance and all members over 60^{th} percentile individual performance
2. Team of Experts: Under 75^{th} percentile team performance despite all members over 60^{th} percentile individual performance.
3. Weak Link: Under 75^{th} percentile team performance with at least one but not all more members under 40^{th} percentile individual performance.
4. Novice: Under 25^{th} percentile team performance and all members under 40^{th} percentile individual performance

3.2 SLATS Diagnostics

While not the main focus of this paper, after a team was classified by SLATS this result could be visualized in a web interface as shown in Figure 2. When providing diagnostic feedback, we consider the generalizable metrics collected and differentiate them by the individual vs. the team [9] and also the team expertise level. Based on the anchor points of Novice and Expert Team as the lowest and highest archetypes, respectively, a rank-order was inferred for the next-better archetype to advance toward.

The team's performance on each feature was shown as a bar chart. A green bar indicates the team's performance on a metric exceeds the typical team in their archetype (red-dotted line) or the next-better archetype (yellow-dotted line). A red bar indicates falling short of the typical standard for the current archetype on that performance feature (e.g., worse than other novice teams). As shown in the third bar "Survive", the next-better archetype might be worse than the current one on certain performance features. Suggested "Sustains" and "Improves" recommendations are displayed below the bar chart. For more expert team, areas to improve will typically be team metrics. However, for more novice teams, areas to improve will more commonly be individual skills to practice.

3.3 Transfer Learning Analysis

To evaluate this approach, TF2 data was used as a proxy for future synthetic battle training. Teams post log files of their matches to public online reposito-

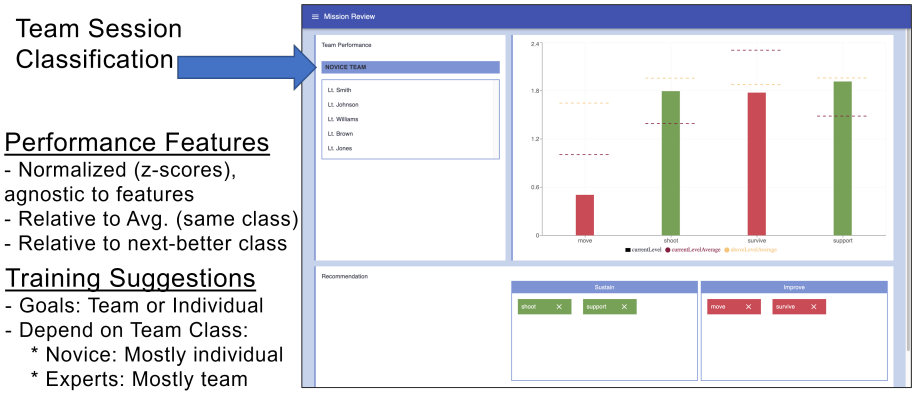

Fig. 2. SLATS Session Diagnostics User Interface

ries and TF2 scenarios require balancing individual competencies (e.g., shooting accuracy, taking cover) with teamwork competencies (e.g., capturing positions, healing/support). We collected a corpus of TF2 matches using the same "mode" (e.g., goal and rules), and classified teams in our corpus of TF2 matches into four archetypes based on the performance of the team and the performance of individual members of the team across the corpus as explained above.

The relationships between team and individual metrics are outlined in Figure 3, with individual metrics in orange and team metrics in blue. The feature vector for a session had four values: Move (capture and hold points), Shoot (kill or damage opponents), Support (heal or assist team member in a kill), and Survive (heal self, avoid damage, and avoid death). The Move metrics included each capture point as a distinct lower-level metric, so that performance was based on the percentage of the session each point was held and the maximum number of points they held at the same time. Team metrics were also normalized to average across the number of players and converting to z-scores for each team session metrics so they would be on comparable scales.

To investigate the feasibility of transfer learning, we isolated matches from two of the most frequent maps/terrains for a game mode called "Control Point" (Snakewater and Process). These matches require capturing and holding a set of control points on the map to win. In each analysis, test data was drawn solely from the target map and we explored the use of varying mixtures of training data from the source and target maps. The SLATS architecture was configured to use default classifier settings (GMM and Logistic Regression), with the expectation that if transfer learning assists simpler models it should also benefit more data-intensive models. A stratified random sample of sessions was selected from each map, which ensured that all archetypes were represented and that only one session per match was selected (i.e., avoiding two sessions from same match but different teams). A total of 319 sessions were processed to generate session feature vectors, with 246 Snakewater (Sn) and 73 Process (Pr) sessions prepared.

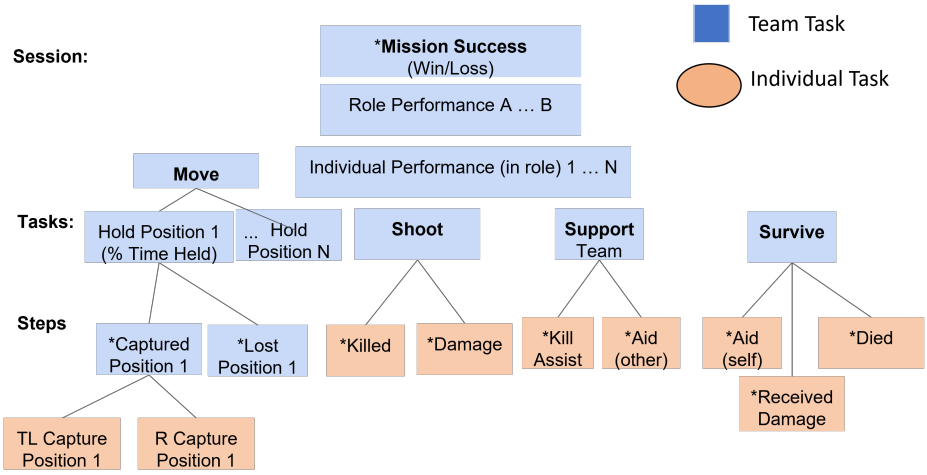

Fig. 3. TF2 Metrics Aggregation Diagram

4 Results

Cross validation on entire session corpus of 319 sessions showed strong classification results (5-fold CV; F1=0.97±0.02). Table 1 shows the average F-1 scores for each additional 14 training sessions from either a source map or the target map (which also provides sessions used as test data).

	Target Training N_{Pr}				
Source Training N_{Sn}	0	14	28	42	56
0	N/A	0.867	0.865	0.895	0.962
14	0.912	0.917	0.945	0.954	0.969
28	0.900	0.959	0.969	0.949	0.987
42	0.936	0.989	0.976	0.967	0.980
56	0.942	0.969	0.980	0.960	0.960
70	0.933	0.939	0.949	0.969	0.987
84	0.962	0.980	0.960	0.987	0.980
94	0.953	0.969	0.966	0.939	0.967

Table 1. F-1 Scores (Avg. of 5-fold CV) for Team Classification based on Source and Target map sessions

Particularly when data is limited, including training samples from both maps improves classification performance on the Target test sessions. These benefits are most pronounced with fewer than 56 Target sessions (F-1 below 0.9 without Source sessions, but 0.912-0.954 with even just 14 Source sessions). A follow-up

analysis with greater randomization of samples confirmed these results, showing that the average best-performance tended to be about F1=.966 and that it typically plateaued at approximately 72 samples (28 Source/42 Target).

5 Discussion

This research found that both semi-supervised learning and transfer learning can improve classification of team performance. As shown in Figure 2 for diagnosis, archetype analysis enabled by the semi-supervised approach is helpful because team performance is not just on a monotonic scale but in cases where some metrics may decrease as teams improve overall. Transfer learning also showed benefits for overcoming the cold start problem of limited data. However, generalizable metrics pipelines were challenging to design when relying on xAPI standards-based approaches, which likely requires more specialized research in this area.

Transfer Effectiveness. Leveraging data from the source map always improved classification f1-scores compared to relying solely upon target map training data. The greatest benefits were observed when target data was limited (0 to 42 target examples). Although it was not always the case that more source data results in higher performance, validation data could be used to find the ideal mixture of source and target training data. While further research is required to explore the effectiveness of transfer learning across training scenarios that are more dissimilar (e.g., different simulations, rather than just different maps), these results offer a promising direction to help bootstrap team assessments on new training scenarios by leveraging data from earlier, comparable scenarios.

Improving Metrics Pipelines. Our work is complementary to research that improves underlying assessment metrics, such as research on multi-modal assessment of training scenarios [11]. Since SLATS archetypes are derived from aligning small amounts of labeled data with larger bottom-up clusters, the specific assessment metric components can be replaced with more advanced measures while following the same pipeline. In addition to more advanced metrics, more efficient calculations of standards-based metrics are also needed. The sheer volume of data for a highly-logged scenario (e.g., TF2) posed challenges in this research. Attempting to apply a standards-based approach for xAPI conversion of each low-level action (e.g., every shot fired) resulted in very large xAPI learning stores. Processing metrics on such records required optimized queries and database caching of results, which undermined the goal of easily generalizing team assessment across different training systems. Research groups have investigated data streams and other techniques to optimize metrics [1], which may offer a foundation for future work on reusable metrics.

Tradeoffs of Archetypes. The SLATS approach depends on interpretable team archetypes, which can be benchmarked against real teams rather than heuristic rules or cutoffs. However, training experts may not know or recognize distinct team archetypes for all training scenarios. In particular, the assumption of an "expert" category assumes that as experts gain skills, they tend to behave

more and more similarly in comparable situations (i.e., converging on the best approach to situations such as landing a damaged aircraft). However, expert teams may also become more diverse in their behavior (e.g., developing their own patterns of communication, developing teamwork patterns unique to the strengths/weaknesses of individuals in the team). Thus, it may be difficult to recognize expert teams in some scenarios because how they coordinate and work together may differ, implying that additional archetypes might be required when these distinctions are relevant for training.

6 Conclusions and Future Directions

Research on the SLATS framework indicates that transfer learning offers advantages for team assessment, particularly when data is limited and relatively simple models are leveraged. Future research is needed however to replicate these findings with more advanced models, particularly models that could incorporate data streams more directly for a high volume of data. For example, new classes of neural network transformer models may be able to directly ingest event data streams and produce meaningful assessments.

These findings suggest that outcomes-based assessment for training scenarios and simulations may someday be automated usefully, not just for assessing team performance but also for tracing individual poor task performance that may need further practice. However, the current work did not model more complex processes or delayed consequences that may occur in other scenarios (e.g., the appropriate skill to practice if a small mistake early-on results in a large failure later). These types of assessments are important, as simulated assessments should distinguish between a good process versus a good (or bad) outcome when suggesting skills to study.

Finally, research on automatically generated formative assessments and diagnoses for training scenarios warrants further pilot studies and evaluation research to indicate how much these insights help guide and improve learning outcomes and study processes. Despite growing interest in automated or partially automated assessment, data on effectiveness remains limited. As such, future work should conduct studies to identify the benefits and limits of such feedback compared to control conditions such as a Wizard-of-Oz model (i.e., feedback controlled by a hidden human expert) or a system without formative assessments.

7 Acknowledgments

This research was sponsored by U.S. Army through the USC ICT University Affiliated Research Center (W911NF-14D0005). However, all statements in this work are the work of the authors alone and do not necessarily reflect the views of sponsors, and no official endorsement should be inferred. This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: `http://dx.doi.`

References

1. Blake-Plock, S., Hoyt, W., Casey, C., Zapata-Rivera, D.: Data analytics and visualization for xAPI learning data: Considerations for a GIFT strategy. In: Design Recommendations for Intelligent Tutoring Systems, Volume 8: Data Visualization. pp. 163–171 (2020)
2. Brawner, K., Sinatra, A.M., Gilbert, S.: Lessons learned creating a team tutoring architecture. In: Design Recommendations for Intelligent Tutoring Systems, Vol. 6 Team Tutoring. pp. 201–220. US Army (2021)
3. Gilbert, S.B., Slavina, A., Dorneich, M.C., Sinatra, A.M., Bonner, D., Johnston, J., Holub, J., MacAllister, A., Winer, E.: Creating a team tutor using GIFT. International Journal of Artificial Intelligence in Education **28**(2) (2018)
4. Goldberg, B., Owens, K., Gupton, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M.: Forging competency and proficiency through the synthetic training environment with an experiential learning for readiness strategy. In: Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (2021)
5. Johnston, J.H.: Team performance and assessment in GIFT – research recommendations based on lessons learned from the squad overmatch research program. In: Proceedings of the 6th Annual GIFT Users Symposium (GIFTSym6). US Army (2018)
6. Neyshabur, B., Sedghi, H., Zhang, C.: What is being transferred in transfer learning? Advances in neural information processing systems **33**, 512–523 (2020)
7. Nye, B.D., Core, M., Auerbach, D., Ghosal, A., Jaiswal, S., Rosenberg, M.: Integrating an engagement classification pipeline into a GIFT cybersecurity module. In: Proceedings of the 8th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym8). pp. 49–67. US Army (2020)
8. Nye, B.D., Core, M.G., Jaiswa, S., Ghosal, A., Auerbach, D.: Acting engaged: Leveraging play persona archetypes for semi-supervised classification of engagement. International Educational Data Mining Society (2021)
9. Smith-Jentsch, K.A., Johnston, J.H., Payne, S.C.: Measuring team-related expertise in complex environments. In: Cannon-Bowers, J.A., Salas, E. (eds.) Making decisions under stress: Implications for individual and team training. American Psychological Association (2018)
10. Triguero, I., García, S., Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowledge and Information systems **42**(2) (2015)
11. Vatral, C., Mohammed, N., Biswas, G., Goldberg, B.S.: GIFT external assessment engine for analyzing individual and team performance for dismounted battle drills. In: Proceedings of the Ninth Annual GIFT Users Symposium (GIFTsym9). pp. 107–127. US Army (2021)