Acting Engaged: Leveraging Play Persona Archetypes for Semi-Supervised Classification of Engagement

Benjamin D. Nye [*]
Inst. for Creative Technologies
Univ. of Southern California
nye@ict.usc.edu

Mark G. Core [*]
Inst. for Creative Technologies
Univ. of Southern California
core@ict.usc.edu

Shikhar Jaiswal [*] [†]
Microsoft Research
Bangalore, India
t-sjaiswal@microsoft.com

Aviroop Ghosal [†]
Amazon.com, Inc.
Detroit, Michigan, USA
aviroopg41@gmail.com

Daniel Auerbach
Inst. for Creative Technologies
Univ. of Southern California
auerbach@ict.usc.edu

ABSTRACT

Engaged and disengaged behaviors have been studied across a variety of educational contexts. However, tools to analyze engagement typically require custom-coding and calibration for a system. This limits engagement detection to systems where experts are available to study patterns and build detectors. This work studies a new approach to classify engagement patterns without expert input, by using a play persona methodology where labeled archetype data is generated by novice testers acting out different engagement patterns in a system. Domain-agnostic task features (e.g., response time to an activity, scores/correctness, task difficulty) are extracted from standardized data logs for both archetype and authentic user sessions. A semi-supervised methodology was used to label engagement; bottom-up clusters were combined with archetype data to build a classifier. This approach was analyzed with a focus on cold-start performance on small samples, using two metrics: consistency with larger full-sample cluster assignments and stability of points staying in the same cluster once assigned. These were compared against a baseline of clustering without an incrementally trained classifier. Findings on a data set from a branching multiple-choice scenario-based tutoring system indicated that approximately 52 unlabeled samples and 51 play-test labeled samples were sufficient to classify holdout sessions at 85% consistency with a full set of 145 unsupervised samples. Additionally, alignment to play persona samples for the full set matched expert labels for clusters. Use-cases and limitations of this approach are discussed.

[*]Denotes equal contribution.

[†]Shikhar Jaiswal and Aviroop Ghosal contributed to this research as student researchers at the University of Southern California.

Keywords

Engagement, Machine Learning, Semi-Supervised, Clustering, Classification, Archetypes, Play Personas

1. INTRODUCTION

Engagement represents a necessary (though not sufficient) condition for learning. Engagement has been shown to impact learning [4] and persistence [9]. Research has also found that engagement is actionable and can be increased [25]. This is a particularly important topic for computer-based learning: unlike in a classroom, where engagement can be assessed and acted on by an instructor in real-time, patterns of engagement are often not visible [10].

However, building engagement analytics for a new system is time consuming. Custom metrics are typically developed and then require substantial data to identify patterns (i.e., the cold-start problem). Worse, the extensive effort to design such analytics is buried in application-specific code. While heuristics are available to infer disengagement, such as response times under 3 seconds [6], applying these to different systems requires benchmarking and calibrating detectors for the content and system. Efforts to analyze engagement often start almost from scratch. This is unfortunate, since research on behavioral engagement has identified patterns which appear to generalize across systems [3, 4, 6, 14, 15].

To address this gap, we are researching a service for analyzing and classifying engagement that relies on a standards-based learning record store [1]. This effort is called the Service for Measurement and Adaptation to Real-Time Engagement (SMART-E). Rather than being optimized to analyze a specific system or data set, SMART-E targets three high-level goals: 1) *Cold-Start Calibration*: ability to identify and benchmark engagement behaviors, which does not require large data sets or in-depth expert analysis; 2) *Re-Usability*: reliance on standards and data available from most learning environments; and 3) *Actionability*: generation of actionable insights, which an instructor or adaptive system could leverage or investigate further.

SMART-E is influenced by two techniques: 1) *semi-supervised learning*, which trains with a small set of labeled data and

a larger set of unlabeled data and 2) *play persona*, behavioral archetypes commonly used for testing and analysis of video games [7, 32]. Our paper describes the process and findings from applying this approach to a data set from a scenario-based tutoring system for training counseling skills. Contributions from this work include a) reviewing features that generalized engagement analytics should consider, b) developing a pipeline for analyzing engagement which does not require expert labeling or application-specific feature engineering, and c) demonstrating the effectiveness of a semi-supervised approach with reasonable data requirements (e.g., about 50 samples each of labeled and unlabeled data) to approximate inferences that experts might make given similar data. As such, this research represents a step toward a generalized framework for diagnosing learner engagement that does not require an expert researcher analyzing data or observing subjects.

2. BACKGROUND AND THEORY

Across the learning science community, engagement is defined and measured in vastly different ways, ranging from split-second physiological responses (e.g., eye tracking, facial affect) to long-term trends lasting months or years (e.g., returning to a system, building social ties) [2, 13]. The research in this paper targets behavioral engagement at the task level (e.g., time spent working through a problem) and session level (e.g., sustained effort to improve performance and learning).

A key reason for this focus is data availability and data interpretability. Most systems collect data logs at these levels and, as described next, substantial research has also identified common behavioral patterns. Research on lower-level affective cues (e.g., facial affect) has found certain actionable events that generalize (e.g., gaze inattention [15]), but other patterns are not trivial to generalize due to differences between individuals or across contexts [28]. Moreover, facial data is often unavailable due to the privacy issues involved with recording learning. Larger time scales are not the focus of this work because engagement levels over those time scales would require longitudinal data and also are more likely to be visible to instructors (e.g., absences).

2.1 Patterns of Behavioral Engagement

Behavioral engagement analysis from log files has shown repeated evidence of useful, actionable patterns, such as response time, response time vs. accuracy/correctness interactions, approach vs. avoidance behaviors relative to problem difficulty (e.g., skipping hard problems), and noisiness of answer quality (e.g., carelessness) [5]. Response time, particularly very fast response time, is one of the most obvious features linked to behaviors associated with disengagement (e.g., guessing, skipping, straight-lining). For scored tasks, the interaction between response time and correctness has been extensively researched in the study of basic cognition as well as authentic learning tasks [6]. The relationship between correctness and time is frequently a logistic relationship (assuming that time does not directly impact scoring): with very fast responses, correctness is approximately random, increasing rapidly to better than chance for more ordinary response time, and approaching an individual skill-level asymptote as time increases. At very large times, answer quality may once again decrease, either due to distraction (e.g., multi-tasking) or difficulty selecting a final answer [27].

More complex interactions often require understanding the relative problem difficulty. Research indicates that students with poor learning outcomes tend to avoid or abuse hints on problems that they find difficult [5]. Conversely, self-regulated learners may be more likely to skip or "game" through problems that that are easier relative to their skill level but dedicate more time to harder problems [33]. While not yet investigated, this might also imply that more self-regulated learners may be less likely to demonstrate wheel-spinning [18] since they are more actively monitoring the usefulness of tasks.

Estimates of answer correctness versus expected correctness have also been used, though these are likely most clear when the learner is close to mastery. Of these, carelessness and "slips" are the most well-established mechanisms [12]. More generally, there may be value in investigating any situation where correctness appears decoupled from traditional factors (e.g., little correlation between time and answer quality, little correlation between expect mastery and later performance). However, such decoupling could be due to poor task design (e.g., item response issues [20]) or problems unrelated to engagement (e.g., attention or memory problems), so additional context may be needed to interpret this.

2.2 Archetypes for Behavioral Engagement

When considering these different patterns of behavioral engagement and disengagement, we posit that engagement has at least two dimensions: a) passiveness vs. activeness and b) avoidance vs. approach. For example, passive avoidance represents disengagement commonly associated with boredom such as distraction or skipping through material. By comparison, other learners employ short-cut strategies to cheat or cherry-pick tasks to minimize effort while still providing acceptable performance (active avoidance). A similar division exists for engaged learners, in that some study almost exclusively on assigned content (passive approach) while others monitor and self-regulate their effort to focus their learning (active approach).

These latent engagement factors may be evident through different observed patterns. For example, while distraction and racing through material both represent disengagement, their data patterns will look very different. In considering these patterns, we developed the following candidates which may be evident across a variety of systems:

- Diligent (Active Engagement): Spends somewhat more time on tasks and shows correspondingly better performance, and more likely to complete optional tasks.
- Self-Regulated (Active Engagement): Seeks out and spends greater time on harder tasks, but may skip or disengage on easier tasks. [22, 33].
- Cherry Picking (Active Disengagement): Seeks out easier tasks or abuses features to make tasks easier (e.g., hint abuse), and avoids harder tasks [3].
- Nominal Engagement (Passive Engagement): Completes tasks as recommended or assigned, with ordinary time-on-task and performance.

- Expert/Recall (Passive Engagement): Regardless of difficulty level, completes tasks very rapidly and with high performance. Possibly an expert on the content, but might also be shallow recall or lookup.

- Racing/Guessing (Passive Disengagement): Rapidly answers (potentially multiple times) despite relatively poor performance [26].

- Distracted/Slow (Passive Disengagement): Uncommonly delayed or irregular answers, particularly when extra time does not appear to improve performance [27].

As with prior research on engagement, we do not assume that these archetypes are necessarily stable for a specific user across all content, but that they represent modes of interaction during learning. Additionally, these candidate patterns are not exhaustive and the specific evidence for each pattern may not be identical: while racing through material might involve rapid guessing in one system, in another it might involve skipping material entirely. Historically, this has meant that detectors are tuned using expert-labeled observations and/or expert feature engineering.

2.3 Play Persona as a Labeling Methodology

This work applies a new approach to generating engagement labels for user sessions. While substantial research has been conducted on engagement, existing methods for determining engagement during computer-based learning are challenging to scale. Our research is intended to complement three methods currently in-use: expert observers, sensor-based affect detection, and self-report [13].

Expert observers can be trained on a specific coding manual until they reach high levels of agreement. Using techniques such as BROMP [29], a trained observer can monitor and label engagement events for multiple students. The primary barriers to collecting this data are the number of trained observers required and issues of privacy and technology (e.g., observing students in online courses). Automated affect detection (e.g, automated facial affect detection) has also been used to analyze engagement [28, 19]. While in principle facial affect scales to a large number of learners, engagement is hard to interpret without also analyzing behavioral patterns (e.g., screen recordings, log files). As with human observers, privacy issues may prevent the necessary recording of data. Moreover, for both human and automated labeling, while learner states may be recorded, they do not include any interpretation about what strategies a learner is using (e.g., focusing on hard vs. easy problems). Self-report offers a different type of engagement label. Users can report their overall engagement and may also be able to describe the learning strategies that they are using [13]. However, self-reported engagement can be affected by reporting bias (e.g., claiming to be more engaged) or subjectivity of engagement ratings.

To address these limitations, we identified play persona as a way to generate labeled engagement data. Play persona are behavioral archetypes often used for testing and analysis of video games, that reflect different goals and behavior patterns [35, 32]. For example, in a strategy game there are recognized archetypes such as the Builder (invest in long-term expansion) versus Greedy-Optimizer (take quick wins) [34]. Likewise, research on Massive Multiplayer Games (e.g., [36]) has identified behavior archetypes such as competitors who focus on head-to-head tasks and explorers who focus on exploring the world. Artificial game players can be crafted to mimic these play persona for procedural play-testing [21].

We hypothesize that play persona methods can also be useful to identify and label engagement patterns with the modification that human testers will act out these roles (e.g., diligent) which would be difficult to simulate artificially. If this approach is useful, it has at least three advantages over existing methods. First, it ensures rapid data collection of labels, since rather than having unbalanced labels (i.e., 80% of real users might be in one bin), testers can be directed to act out a variety of roles. Second, play-test labels should be interpretable since the intent of the learner is known, as opposed to purely bottom-up patterns or self-reported labels, which require experts to infer underlying strategies. Finally, despite some constraints (e.g., difficulty in faking more or less knowledge), dedicated testers may be able to play out multiple archetypes and do so repeatedly, reducing the need to recruit new testers.

3. RESEARCH QUESTIONS

This work investigates techniques to leverage play-testing data for detecting engagement patterns. However, this approach will only be feasible if testers reasonably approximate the behavior of real users. It also relies on the assumption that while systems may differ, the main engagement archetypes will be fairly predictable (e.g., some users will be highly invested in learning every piece of content, others will be trying to get through as fast as possible). In this work, we examine the feasibility of play-testing to help classify engagement patterns, and in particular investigate the following questions:

- Q1 (Distinctiveness): Are the data patterns for a set of play-tester archetypes distinct (different testers act similarly, given similar instructions)?

- Q2 (Alignment): Will play-test archetypes align with unsupervised clusters producing labeled clusters similar to how experts would label them?

- Q3 (Semi-Supervised Comparison): Will a semi-supervised approach that builds a classifier from play-test and aligned data label individual learners more consistently than relying only on bottom-up clusters?

- Q4 (Basic Features): Will average response time and scores, in simple systems, be sufficient for reasonable engagement labels?

- Q5 (Expanding Features): Will increasing the number of features to include task difficulty and feature interactions lead to greater consistency in fewer samples?

These questions investigate the strengths and limitations of the approach. Specifically, Q1 and Q2 focus on the reliability of play-test labels to label unsupervised data, as compared to human ratings. Q3 examines if building a semi-supervised

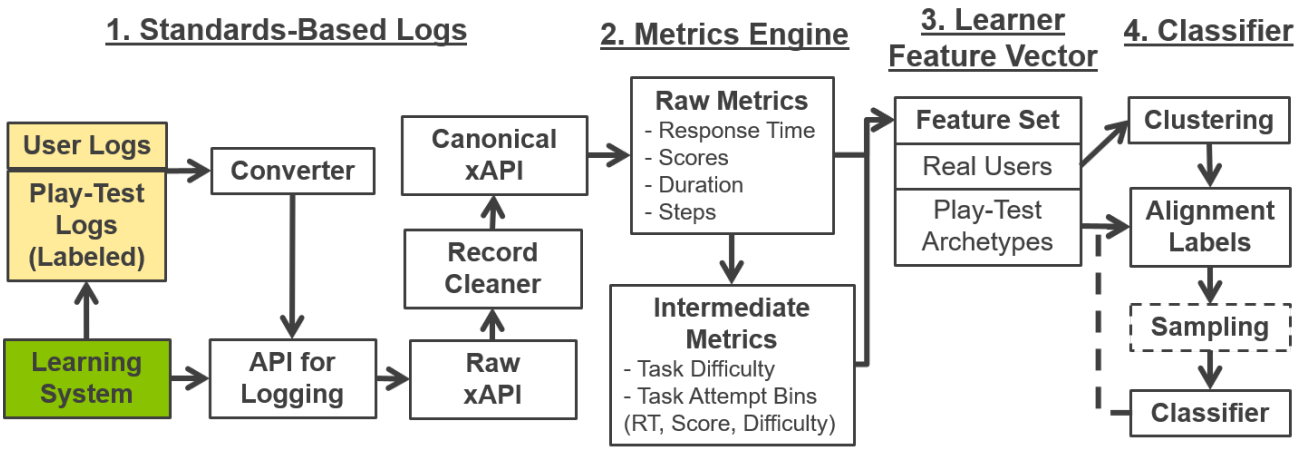


Figure 1: SMART-E Analytics Pipeline Phases

classifier is useful as opposed to simply using archetypes to label bottom-up clusters. Q4 and Q5 query the effectiveness of feature sets identified in the literature for classifying engagement, starting with a very minimal set (response time and scores) and then analyzing an expanded set of features for their impact on cold-start performance.

4. METHODS

To examine these questions, an analytics pipeline was developed and then applied to a data set from a scenario-based intelligent tutoring system. This section will briefly describe the pipeline, then the learning system that produced the data set, and finally the techniques used to investigate each research question.

4.1 Engagement Analytics Pipeline

While this paper focuses on a specific data set, the techniques applied here are designed to be generalizable and reusable as part of the SMART-E pipeline, shown in Fig. 1. This pipeline starts by standardizing the data available, recording data from an arbitrary learning system as learning records that meet the xAPI standard [1]. This "Raw xAPI" data may either be sent directly by the system (e.g., through an API for logging) or generated by running a converter on system logs after-the-fact. Raw xAPI data logs are then cleaned by a script (partially system-specific) which corrects common data problems, such as sessions that terminated improperly or missing data fields that can be inferred from other data. This ensures that the Canonical xAPI data store does not have missing data.

All xAPI records contain metadata which allow them to be structured into an activity tree, representing both sequential and parallel tasks. While the tree can be nested arbitrarily, four levels are analyzed to generate raw metrics tables: steps, tasks, lessons, and sessions. Raw metrics primarily record time-based information (e.g., duration of a task, response time for first step), score-based information (e.g., numerical score and/or correctness), and support used (e.g., hint counts, retrying a problem).

Metrics related to task skills are not calculated, since the majority of systems do not tag their tasks with a consistent ontology of knowledge components. Intermediate metrics are generated using feature construction calculations based on raw metrics, without analyzing the xAPI logs. For this work, the most important intermediate metrics are averages across attempts (e.g., average scores, average task duration), the average difficulty for each task (inferred from first-attempt scores), z-scores for task metrics (e.g., time-on-task for the learner relative to other users) and a Laplace-smoothed logarithm of each task duration (i.e., $ln(t+1)$). Additional metrics can be added fairly rapidly, if they rely on raw metrics.

Based on these metrics, feature vectors are generated that represent each learner's performance in the system. In the current work, these vectors rely on all of the learner task data for a session, though one could generate similar features for specific tasks, across multiple sessions, or for recent tasks in a session (i.e., any collection of tasks). First, two simple features were calculated: average response time across tasks (Avg. RT) and average task performance (Avg. Score). These were considered the minimal information to potentially infer engagement.

Next, a more complex feature set was developed to model interactions between task response time (RT), task scores, and task difficulty. Based on z-score cutoffs, the value of each variable was placed into one of three categories (low, medium or high) when possible, and into the most categories available when not (i.e., only medium if all values equal; only low and high if only two types of values). This was done based on a one-dimensional Gaussian distribution, with cutoff values at <33% (low), 33-66% (medium), and >66% (high). Further we ensured that each variable had at least 4 corresponding data-points in order to arrive at robust cutoffs (i.e., each unique task had been attempted by at least 4 different learners, to judge its difficulty, score and time distribution). Each scored task increments a bin associated with its three variables (e.g., RT=fast, score=high, difficulty=high will increment exactly one out of 27 possible bins). This binning approach is fairly general, and can be inferred using only standard logging data.

Since 27 bins will often be fairly sparse for an individual

learner, these were aggregated to form 7 bins which align to behavioral engagement patterns from the literature: Expert, Cherry Picking, Engagement/Diligent, Self-Regulated, Distracted, Racing, and Careless. These bins roughly correspond to the patterns we introduced earlier except we omitted Nominal Engagement, roughly equivalent to average, and we added a Careless bin focused on errors on easier tasks. The Expert bin was increased whenever high scores were obtained for difficult problems with only a normal or low delay or for high scores on ordinary problems done quickly. The Cherry Picking bin incremented for high scores with a low delay (regardless of problem difficulty). Engagement/Diligent was incremented when difficult problems were completed after a high delay. Self-Regulated was incremented when the amount of time spent on the problem was at least as high as the difficulty level, even if the score was not high. Distracted was triggered in the opposite case, where the time to respond was overly long for the difficulty of the problem. Racing was incremented for fast responses, either with low scores or with medium / high scores on easier problems. The Careless bin included only low scores on easy problems or low scores on medium difficulty problems when completing them quickly.

These bins were not mutually exclusive, since more than one behavior might explain a given interaction. Additionally, they are not validated and should be thought of as noisy constructs to bin low-level features, rather than necessarily predictive of their given labels. However, since these aggregation patterns are derived from the literature, these features are candidates that may be relevant across different systems, users, or data sets.

4.2 User Data: ELITE Scenarios

We use data from the system, ELITE Lite Counseling, designed for U.S. Army officers in training to learn leadership counseling skills, such as active listening, checking for underlying causes, and responding with a course of action [11]. Learners select what to say to virtual subordinates from a menu leading to different points in a branching graph representing the possible conversations. The virtual subordinates speak using pre-recorded audio and act via 3D animations.

Each learner choice can have both positive and negative annotations. Positive annotations correspond to correctly applying a skill such as active listening, and negative annotations correspond to omissions or misconceptions. Based on these annotations, a choice can be fully correct (only positive annotations) or two forms of incorrect: fully incorrect (only negative annotations) or mixed (both positive and negative annotations). For the pipeline, this was converted to two forms: a correctness category and a numerical score in which mixed answers were given partial credit (0.5) compared to correct answers (1.0) and incorrect answers (0.0).

Each simulated conversation is also followed by an After Action Review (AAR) in which learners are asked multiple-choice questions about all of their dialogue choices that were mixed or incorrect. For these AAR questions, if the first attempt to answer was successful the learner earned a score of 1; otherwise, the learner earned a score of 0 but had to keep trying until they selected the correct response.

The ELITE data set for this research included a corpus of 145 subjects from experiments described here [17] which we consider user data. Each "user" completed three scenarios: Scenario 1, Scenario 1 (Repeated), and Scenario 2. Due to the dialog trees, users did not all see the same decision tasks when completing the same scenarios. However, substantial overlap was observed for tasks and a majority of tasks were attempted by a significant number of users. For the purpose of estimating task difficulty, a threshold of 5 attempts was used, below which the difficulty and metrics relying on difficulty (e.g., binning) could not be calculated.

4.3 Play Persona Data

Play-testers were students and employees of the lab who volunteered their time, and generally were not familiar with the scenario content. For the ELITE system, only 5 play-test archetypes were reasonable to classify: Expert, Diligent, Nominal Engagement, Racing, and Distracted. Play-testers followed the same protocol (i.e., scenario 1, scenario 1 repeated, scenario 2) used to collect the user data. Thus, they had no direct control over the tasks they encountered, and so some patterns were unlikely to be observed (e.g., Self-Regulated and Cherry Picking).

Each play-tester was able to generate data for up to three archetypes, by attempting them in a specific sequence. First, they could play as either Diligent or Distracted. These roles could only be played at the beginning of testing to simulate a novice seeing the system for the first time. Next, a Racing run was completed; fast response times meant testers would still make errors despite their previous practice. Expert runs were collected in two ways: either an actual expert generated the data (2 sessions), or a tester carefully reviewed the correct answers (e.g., in the AAR) and/or was coached by an expert (13 sessions). These different methods for "expert" data produced similar results, though actual experts were slightly faster. In the unlabeled data, an archetype for Nominal Engagement was generated by extracting five clusters and assigning Nominal Engagement to the cluster not aligning with the other four archetypes.

Instructions for each play-test archetype were as follows. **Diligent:** spend as much time as you need on each choice to try to get the best answer, including reading carefully, and double-checking answers. **Distracted:** engage in one or more competing activities, including checking email and responding when relevant, browsing social media, engaging in a conversation, and eating. **Racing:** pretend you don't care much about the content, so you are doing the bare minimum and are fine with a so-so score to get done quickly. **Expert:** review content in-depth immediately ahead of time, and approach it with as many answers memorized or quickly-available as possible (e.g., in notes, from an expert) so you can answer well quickly. Of these, all except Distracted were easily understood by testers. Due to the lack of standardization for Distracted, some testers struggled to find a competing distraction task (e.g., did not use much social media, did not have high email volume, already ate lunch). In this case, a member of the research team asked the user questions or other requests to distract them. A total of 51 archetype sessions were collected, which may be more than necessary, since preliminary analyses found similar results with about 25 points balanced across classes.

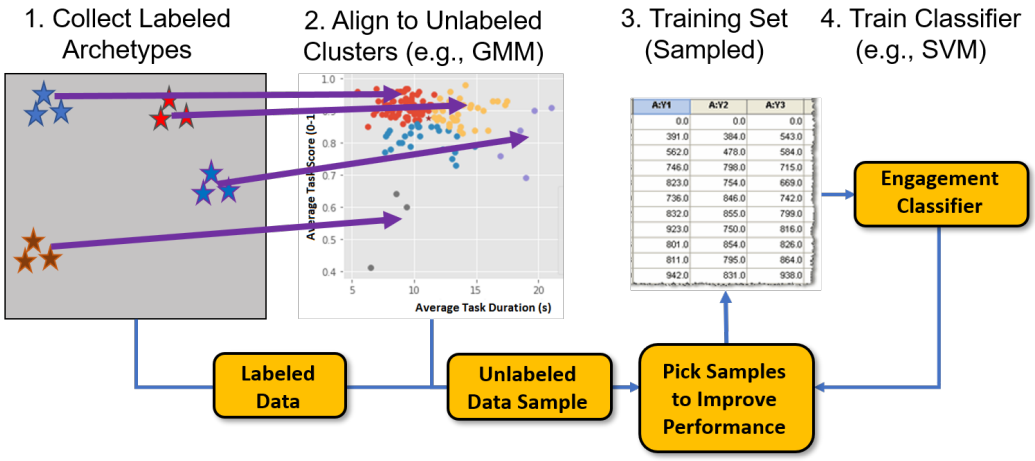


Figure 2: Semi-Supervised Classifier Training

4.4 Cluster Alignment Testing

As shown earlier in Fig. 1, both the real user data and the play-tester archetype data were processed by SMART-E to generate feature vectors that represent each individual. A number of techniques were then applied to generate labeled clusters. This cluster-labelling process allowed us to classify a learners' engagement coarsely on the basis of the cluster that they were assigned to.

First, user feature data was clustered bottom-up into five distinct clusters using k-means and Gaussian mixture models (GMM) methodologies as implemented in the scikit-learn package [30]. The number of clusters was verified through an elbow-curve analysis of variance explained (elbow at k=5). Exploration of k=4 and k=6 found both to be less stable; cluster assignments were often very different for subsamples of data points, with k=4 being particularly unstable.

For this analysis of the full sample, we associated each user cluster with a unique archetype (i.e., alignment of the smaller archetype clusters with the user clusters). The alignment was determined using the Hungarian Method (Kuhn-Munkres algorithm)[24], which is a global, optimal-matching algorithm which minimized the sum of the Euclidean distances between these user cluster centroids and archetype centroids. As noted previously, the Nominal cluster was determined as the cluster remaining after all archetype groups were matched. As a result, each of the user clusters (and consequently the points within that cluster) had an associated unique archetype which additionally served as its label. When this cluster alignment process is used as the only technique to label points, it will be referred to as *Clustering Alone*.

4.5 Semi-Supervised Classification

This technique of clustering alignment was compared against a semi-supervised approach that built a classifier using the play-test and user clusters. The high level concept of this semi-supervised classifier is shown in Fig. 2. The first two steps of the semi-supervised approach are the same as Clustering Alone. This generates a pool of weakly-labeled candidate labeled points. The points in this pool can be either taken as a full set to train a classifier model such as SVM or they can be sampled to incrementally train a classifier using active learning techniques until a stopping rule is hit (e.g., entropy sampling).

To compare the classifier against cluster-level labels based on archetype alignment alone, we calculated two quality metrics for the labels given to user sessions, which we will term *consistency* and *stickiness*. Consistency refers to the fraction of sessions that are labeled with the same engagement archetype which they would receive when the full data set is available. This is important because as data gets larger, unsupervised clusters are more likely to reflect the true distributions.

Stickiness refers to the likelihood that a user session retains the same engagement label after a batch of new data is added (similar to intra-rater reliability). This is important for actionable engagement metrics: if Student A is classified as Diligent, it will be confusing if Student B who completes an identical run is classified differently due to data that arrived in between. While this cannot be fully avoided, approaches that tend to keep the same label for an identical session will appear more fair and reliable, so that an instructor could be more confident in using the classifications.

That said, neither consistency or stickiness alone are sufficient for useful classification. For example, always assigning all users to the same category maximizes both metrics. However, assuming clusters for the full data set are reliable, then these measures help to identify how quickly and reliably labels approximate the final labels. This is important for addressing the cold start problem, so that engagement patterns can be quickly identified in a new system.

To calculate the number of samples to reach a given level of cold start performance, random splits were made of the user data set into train-test subsets (115 train, 30 test). For each random split, the classifier was trained using the archetype data set (51 samples) and increasingly larger subsets of the user training data in increments of 5. When evaluating cold-

start performance, a consistency of 85% was considered a reasonable target threshold for reliability against the full sample. While the actual consistency required will depend on the specific application, this cutoff should give some insight into how quickly different approaches converge toward their larger-sample performance.

Since the pipeline parameterizes the specific algorithms, follow-up exploratory analyses were conducted with different types of clustering algorithms (e.g., k-means, GMM), classification algorithms (e.g., logistic regression, support vector machines), and semi-supervised sampling algorithms (e.g., full sampling, margin sampling with stopping rules to exclude certain unsupervised samples). Different combinations of these algorithms did not show qualitatively different end-results on these metrics, and any differences were not conclusive (e.g., GMM clusters appeared slightly more stable than k-means as data was added, but within random variation). As a result, this paper presents results for the GMM clustering with a Support Vector Classifier, where these results are representative for the different approaches explored.

5. RESULTS

Focusing on GMM clustering, we revisit the alignment of the five user clusters with the play-tester archetype groups. The clusters were generated using the average of the logarithms for task response time (Log-RT) and average of task scores (Scores). Fig. 3 plots the real user data with unsupervised clusters. Table 1 shows feature means and standard deviations for each archetype, above its most closely-aligned bottom-up cluster. Note that while Log-RT was used for clustering, the actual time in seconds is given in the table and figure for easier interpretation.

Despite being generated independently, the play-test data closely resembles the real bottom-up clusters. As a trend, the play-tester archetypes tend to be more extreme (i.e., farther from the average user) than the clusters they align to. This is likely due to play-testers acting out more exaggerated or consistent patterns than real users. However, this may actually be an advantage, since play-test archetype data points may be more likely to be outliers in the vector space and good anchors for distinct clusters. The results from Table 1 support research question Q1, in that play-testers were able to act out similar patterns as real users and that the play-tester data showed fairly distinct groupings (as evident in the standard deviation values). One exception was the Distracted archetype, which had a very high variance for time compared to real users in the corresponding cluster. However, despite the high variance, the Distracted archetype data remained distinct from other archetypes' data.

5.1 Reliability vs. Expert Labels

The validity of this alignment on the full data set was evaluated by surveying a set of external engagement experts (N=5) to label the same bottom-up clusters obtained from the user feature data, based on the descriptions of the engagement archetypes. Selection criteria for experts required a Ph.D. in a relevant area, publishing at least one substantial paper researching learner engagement, and having no prior experience with the data set.

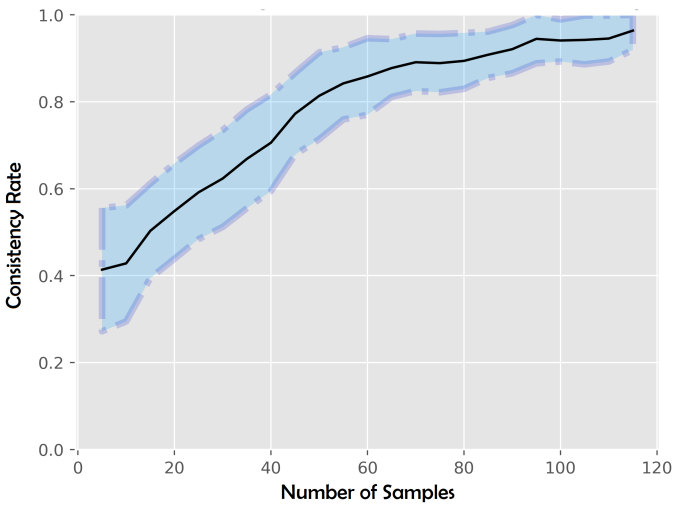
Experts labeled cluster graphs (e.g., Fig. 3) generated by both k-means and GMM, and maintained quite similar labels across each (76% agreement). Since the clusters and labels for both GMM and k-means were very similar, all labels were treated as examples from the same task. Inter-rater reliability metrics were moderate between experts: 55% Agreement; Fleiss' kappa = 0.44; Krippendorff's alpha = 0.45. Expert raters had very high reliability for Expert and Racing labels, but approximately half of experts demonstrated a consistently different interpretation for Diligent, Nominal (phrased as "Average" in the survey), and Distracted. Based on open response comments, this may have been the result of interpreting minor wording differences in the prompts (e.g., "novice learners" for Diligent vs. "learners" in Distracted).

The human labels for clusters were then compared pairwise against the automated alignments, resulting in Agreement, Fleiss' Kappa and Krippendorff's alpha metrics which were higher than within-experts though still in the moderate range: 66% Agreement; Fleiss' kappa = 0.57; Krippendorff's alpha = 0.58. Given these results and expert sensitivity to the wording of archetype descriptions, we conclude that the automatic alignment appears to be at least as useful as expert consensus ratings for labeling engagement clusters. We anticipate automatic alignment to be even more advantageous when the feature space expands beyond 3 dimensions, making it difficult for human experts to visualize or evaluate.

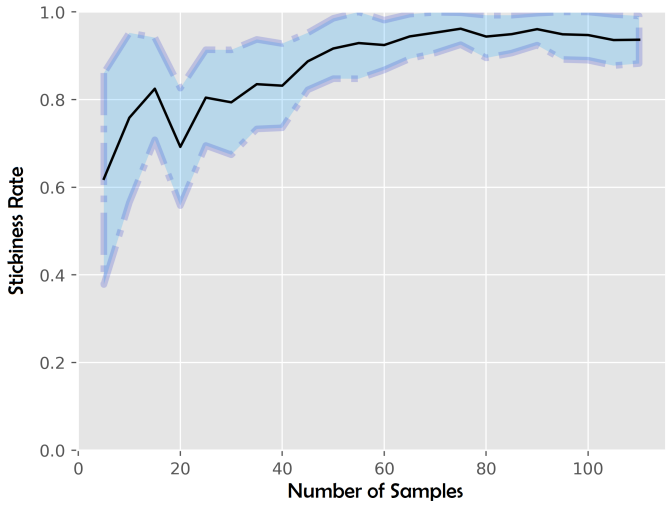
Group	N	Avg. RT (s)	Avg. Score
Expert (Arch)	15	8.53 ± 2.43	0.95 ± 0.04
Cluster 1	25	8.10 ± 1.00	0.93 ± 0.03
Diligent (Arch)	14	13.15 ± 3.83	0.89 ± 0.07
Cluster 2	75	11.06 ± 1.61	0.90 ± 0.03
Nominal (Arch)	-	-	-
Cluster 3	13	8.63 ± 1.11	0.82 ± 0.02
Distracted (Arch)	12	22.27 ± 13.80	0.77 ± 0.17
Cluster 4	28	15.81 ± 3.43	0.83 ± 0.07
Racing (Arch)	10	7.18 ± 2.47	0.56 ± 0.17
Cluster 5	4	7.98 ± 1.08	0.55 ± 0.09

Table 1: Cluster vs. Archetype Centers ($\mu \pm \sigma$)

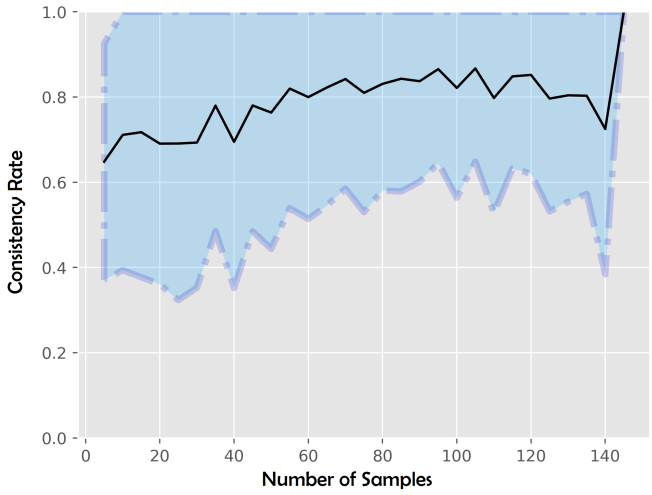
Figure 3: GMM User Clusters for Response Time and Score Features



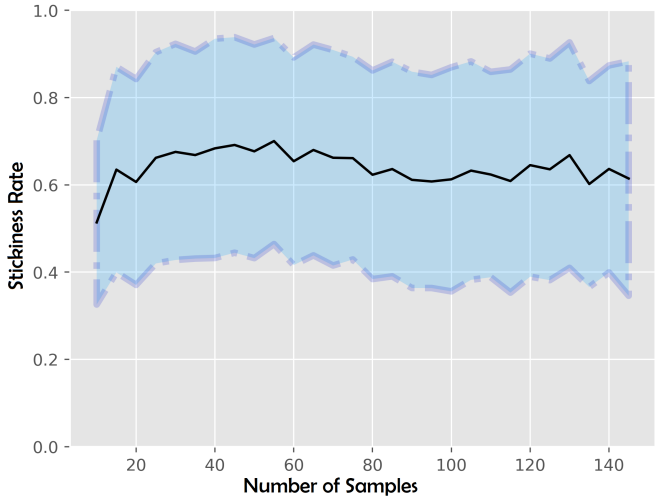
(a) Consistency: Semi-Supervised SVM



(b) Label Stickiness: Semi-Supervised SVM



(c) Consistency: Clustering Alone



(d) Label Stickiness: Clustering Alone

Figure 4: Consistency and Stickiness for Semi-Supervised vs. Clustering Alone

5.2 Consistency of Semi-Supervised vs. Clustering Alone

To evaluate how play-test data can be used to classify new user sessions, a semi-supervised approach was explored which trained a Support Vector Machine (SVM) classifier using both the play-test archetype data and the data from the bottom-up cluster that best aligned to each archetype, with test-set labels determined by the classifier. For the clustering alone comparison case, bottom-up clusters were directly aligned against archetype data to determine their labels and test-set labels were determined based on their closest cluster. 20 random splits were made of the user data set into train-test subsets (115 train, 30 test). For each random split, the classifier was trained using the archetype data set (51 samples) and increasingly larger subsets of the training data set in increments of 5, and then evaluated.

Consistency was calculated against the test set of 30 samples. Stickiness of labels was calculated for each set of labels against the prior set (e.g., model trained on N samples vs. N-5 samples). Due to the higher level of noise for clustering alignment alone, 100 runs were conducted instead of 20 for a smoother average. These results indicated that training a classifier which combined both types of data produced higher consistency and less variation. Specifically, on the basic features (avg. RT and performance only), the semi-supervised SVM reached 85% average consistency at 52 samples (Fig. 4a), while aligned clusters alone required 95 samples to reach this level (Fig. 4c). Clustering alone was more consistent with the full-data cluster labels until approximately 25 samples (i.e., when the user data reached approximately half of the archetype data).

Likewise, the stickiness of labels as data increased reached an average of 85% by 45 samples for the semi-supervised classifier (Fig. 4b). Clustering alone never reached 85% and remained less than 70% on average (Fig. 4d). For both metrics, the variance (blue bars) were larger for clustering alone. One reason for greater variability for clustering alone

is that sparse data for certain cluster regions (e.g., Racing, with only 4 real users), so alignment alone may try to align a non-existent cluster given limited data. However, the semi-supervised classifier appears to mitigate this issue since training is anchored by play-test data points.

These analyses were performed using both the basic features and the expanded feature set (e.g., bins that count instances of engagement behavior patterns based on response time, score, and difficulty categories). Both feature sets required a similar number of samples to reach the same level of consistency (e.g., about 85% consistency after 50 samples). While it is possible that the expanded feature set might produce more valid labels for an instructor (e.g., better reflecting the categories of users who an instructor might follow-up with), this will not be due to improved cold-start performance.

5.3 Semi-Supervised Class-wise Consistency

An analysis of the consistency for labels in individual clusters (Fig. 5) shows similar insights to the overall clustering label consistency. Points in larger clusters (e.g., Diligent, Expert) are consistent fairly quickly. However, small clusters (Racing) may have few/zero examples even when considering as many as 40 data points, and even with 100 data points have poor consistency. As such, classes with few examples might only be useful for a smaller set of use-cases (e.g., sufficient to share with an instructor, but possibly not reliable enough to take an automated action confidently). We also note that based upon the stickiness analysis (Fig. 4d), performance may be limited by the instability of clustering (points moving between clusters even with nearly full data).

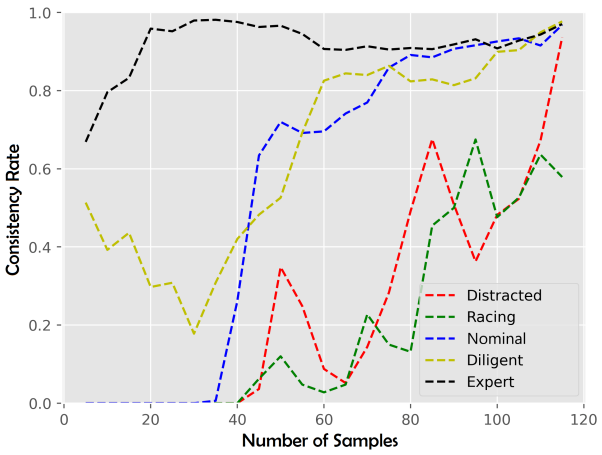


Figure 5: Class-wise Consistency for Semi-Supervised SVM

5.4 Semi-Supervised vs. Final Clusters

The semi-supervised results were compared for their agreement with the labels obtained via alignment with the final clusters generated using the full data set. This final-clusters reference point (see Fig. 3) was used to calculate average accuracy, precision, recall and F-scores (Fig. 6), as a function of the increasing dataset size. While final clusters are not a perfect reference, it shows that accuracy versus final clusters increases fairly rapidly, but that precision, recall and F-scores are consistently lower.

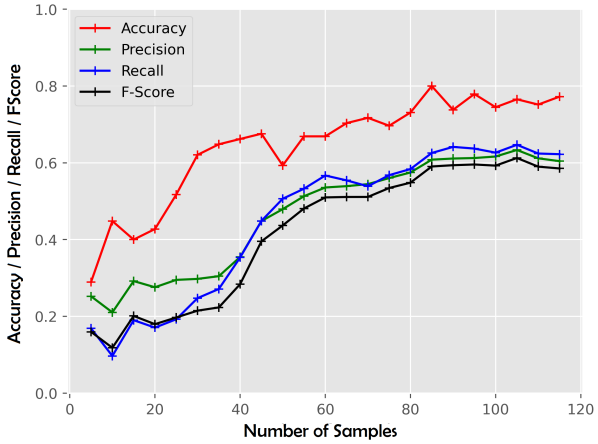


Figure 6: Agreement of Semi-Supervised SVM vs. Final Cluster Labels

6. DISCUSSION

Based on the results presented, this work demonstrates the feasibility of using a play-testing methodology for detecting behavioral patterns of engagement. Moreover, this work also found that a classifier could be developed using this approach without engineering application-specific features. The classifier also offered reasonable cold-start performance and labeled engagement data fairly consistently for 5 categories after 52 unlabeled samples and 51 archetype samples.

Of the five research questions investigated, there was positive support for four answers, with one left indeterminate. For Q1, play-tester data was distinctive and archetype data followed coherent patterns on features (e.g., response time, correctness). Archetype data did not show substantial overlap between archetypes, even though play-testers received only limited instructions. This may be due to the limited degrees of freedom for the task. In a more complex or open-ended system, increased variation might lead to less coherent archetype data. With that said, many systems have similar characteristics to the ELITE scenarios studied here (i.e., sequential linear or branching choice tasks, mixed with passive content such as videos or animations). Moreover, these kinds of systems are often problematic for engagement, such as mandatory corporate training modules.

For Q2, it was demonstrated that automated matching of play-test archetypes against pre-defined clusters performed comparably to expert labels for the same clusters. While refining instructions might improve inter-rater reliability on this specific task, the features presented to experts were already chosen to be simple and visualizable so this represented an optimistic scenario for expert cluster labeling. On more complex feature sets or systems, expert analysis might not even be possible. The broader question not explored in this work is the machine vs. human play-test agreement if they were not given pre-defined clusters (i.e., a data exploration task). However, this would be challenging to conduct: it requires a deep analysis by each expert researcher and the types of engagement categories might be highly uneven. Alternatively, archetypes might be determined from already-analyzed data sets (e.g., such as for hint-abuse), to

see how effectively traces of play-test disengagement might match authentic disengagement patterns.

For Q3, it was established that training a classifier with both play-test data and unsupervised cluster data showed advantages over simply re-clustering with new unsupervised data and then aligning clusters to archetype data. In some respects, this is not surprising: while the consistency metric used for evaluation is based on the unsupervised results from the full data set, the classifier is able to train with more data up-front (as much as double initially). More importantly, since all key archetypes are present in the play-test data, no category will start unrepresented. This particularly helps for classifying points from relatively rare but distinctive categories (e.g., Racing). However, despite this advantage, points in small classes remained substantially less consistent than those in larger classes.

As a long term issue, it is an open question about the best way to mix this data. Neither data source represents ground truth. The archetype data demonstrates coherent engagement patterns, but these patterns might not reflect the ways real users experience the system (e.g., in the current research, they were exaggerated/overly extreme). The real user data is authentic, but may slowly wash out the classifier with unremarkable samples (e.g., overly ordinary). Exploratory work was conducted where stopping rules were applied to balance the number of archetype vs. authentic samples (derived from active learning techniques, such as margin sampling and entropy sampling), but this has not yet produced obvious improvements. Similarly, techniques for weighting samples might be applied. However, the ideal balance between these data sources probably depends on the target use-case for the classifier. A recommender system may want a classifier that acts on labels regardless of their confidence scores. By comparison, a human instructor might prefer a narrowly-scoped but highly-actionable classifier, which might detect clear outliers but allow the majority of user sessions to be in a non-descript "Nominal" category or not confidently classified.

On questions about the features required to classify engagement, we found that basic features for the log of response times and scores were sufficient in this case (Q4) but did not show improvement with the expanded feature set including task difficulty and feature interactions improved classification (Q5). These features helped to detect engagement behavior that matched patterns observed from play-testing: Expert/Recall, Diligent, Racing, and Distracted as well as Nominal (i.e., matched by exclusion). However, both k-means and GMM tended to split up the mass of points in the region of Expert, Diligent, and Nominal despite these clusters being adjacent to each other. The cluster-alignment approach used in this work was selected primarily for the ability to interpret cold-start trends, while more advanced methods should further improve performance. It might be preferred to investigate techniques such as anomaly detection, which would favor a larger central cluster and smaller outliers which could correspond to atypical behavior which is actionable. Alternatively, alternate semi-supervised techniques are available, such as applying specialized semi-supervised support vector machines (which optimize margins for both labeled and unlabeled data) [8, 31] or more advanced techniques for integrating cluster data [16]. While expanded features did not improve consistency or stickiness metrics (Q5), other systems may still benefit from expanded features. However, additional features also increase the required data and may result in overfitting, need attenuating/filtering features during clustering, or other trade-offs. As such, further research is needed on this problem.

7. CONCLUSIONS AND FUTURE WORK

Based on these findings, this work contributes a number of novel approaches to analyzing engagement. First, this research demonstrates the utility of play persona data gathered during professional or quality assurance testing for training useful data mining algorithms. Since there is no definitive metric for engagement, play-test data offers an additional distinct data source to help recognize engagement and disengagement. To our knowledge, this approach has not been applied to analyzing engagement in learning.

Second, this approach offers advantages over current approaches for cold-start labels. Since the behavioral intentions of the play-test users is known with confidence, these labels offer a good data set to help overcome cold start problems. As compared to traditional approaches such as training observers or collecting in-the-moment self-reported engagement [13, 29], play persona data can be collected prior to real system users. This approach also allows balanced sampling for important but lower-frequency engagement behaviors (such as racing, in this analysis).

Third, we have demonstrated that semi-supervised classifiers trained based on a combination of play-test labels and unlabeled data offer more consistent labels than relying on clustering alone, which has been used to analyze engagement behaviors [23]. Moreover, as shown by agreement with expert labels at the cluster level, the alignment approach can provide similar insights without manually interpreting clusters. While expert interpretation is still ideal, this allows immediate insights without waiting for an expert analysis.

This approach is also pragmatic: System developers should already test and perform quality assurance on their software and content [35]. Behavioral archetype data can be collected during this process, by having testers play out engagement styles in a prescribed order based on their expected learning. Moreover, this work is not unique to specific archetypes: if learners are expected to engage in different patterns, play-testers may be able to produce those patterns instead. However, not all archetypes may be realistically playable by testers. For example, experts cannot typically generate novice answers. As such, this approach may be most effective when testers are similar to authentic users. As such, future work will explore how expert observer labels and self-report data might complement this play persona data.

8. ACKNOWLEDGMENTS

This research was sponsored by U.S. Army through the USC ICT University Affiliated Research Center (W911NF-14D-0005). 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.

9. REFERENCES

[1] Advanced Distributed Learning. *xAPI Specification*, 2020.

[2] R. D. Axelson and A. Flick. Defining student engagement. *Change: The magazine of higher learning*, 43(1):38–43, 2010.

[3] R. S. Baker, A. T. Corbett, K. R. Koedinger, S. Evenson, I. Roll, A. Z. Wagner, M. Naim, J. Raspat, D. J. Baker, and J. E. Beck. Adapting to when students game an intelligent tutoring system. In *International Conference on Intelligent tutoring systems (ITS)*, pages 392–401. Springer, 2006.

[4] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.

[5] R. S. Baker and L. M. Rossi. Assessing the disengaged behaviors of learners. *Design recommendations for intelligent tutoring systems*, 1:153–163, 2013.

[6] J. E. Beck. Engagement tracing: using response times to model student disengagement. In *International Conference on Artificial intelligence in Education (AIED)*, pages 88–95. IOS Press, 2005.

[7] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[8] O. Chapelle, V. Sindhwani, and S. S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9(Feb):203–233, 2008.

[9] S. L. Christenson, A. L. Reschly, and C. Wylie, editors. *Handbook of Research on Student Engagement*. Springer, New York, 2012.

[10] M. Cocea and S. Weibelzahl. Disengagement detection in online learning: Validation studies and perspectives. *IEEE transactions on learning technologies*, 4(2):114–124, 2010.

[11] M. G. Core, K. Georgila, B. D. Nye, D. Auerbach, Z. F. Liu, and R. DiNinni. Learning, adaptive support, student traits, and engagement in scenario-based learning. In *Proc. of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, 2016.

[12] R. S. d Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *International conference on intelligent tutoring systems*, pages 406–415. Springer, 2008.

[13] M. A. A. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1, 2019.

[14] S. D'Mello and A. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[15] S. D'Mello, A. Olney, C. Williams, and P. Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398, 2012.

[16] H. Gan, N. Sang, R. Huang, X. Tong, and Z. Dan. Using clustering analysis to improve semi-supervised classification. *Neurocomputing*, 101:290–298, 2013.

[17] K. Georgila, M. G. Core, B. D. Nye, S. Karumbaiah, D. Auerbach, and M. Ram. Using Reinforcement Learning to Optimize the Policies of an Intelligent Tutoring System for Interpersonal Skills Training. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2019.

[18] Y. Gong and J. E. Beck. Towards detecting wheel-spinning: Future failure in mastery learning. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 67–74, 2015.

[19] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *International Conference on Multimodal Interaction (ICMI)*, pages 42–49. ACM, 2014.

[20] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of item response theory*. Sage, 1991.

[21] C. Holmgård, A. Liapis, J. Togelius, and G. N. Yannakakis. Evolving models of player decision making: Personas versus clones. *Entertainment Computing*, 16:95–104, 2016.

[22] R. Janning, C. Schatten, and L. Schmidt-Thieme. Perceived task-difficulty recognition from log-file information for the use in adaptive intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(3):855–876, 2016.

[23] M. Khalil and M. Ebner. Clustering patterns of engagement in massive open online courses (moocs): the use of learning analytics to reveal student categories. *Journal of computing in higher education*, 29(1):114–132, 2017.

[24] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[25] B. Lehman, S. K. D'Mello, A. C. Strain, M. Gross, A. Dobbins, P. Wallace, K. Millis, and A. C. Graesser. Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In *International Conference on Artificial Intelligence in Education (AIED)*, pages 171–178, 2011.

[26] D. J. Leiner. Too fast, too straight, too weird: Post hoc identification of meaningless data in internet surveys. *SSRN Electronic Journal*, 2013.

[27] E. Mattheiss, M. Kickmeier-Rust, C. Steiner, and D. Albert. Approaches to detect discouraged learners: Assessment of motivation in educational computer games. *Proceedings of eLearning Baltics (eLBa)*, 10:1–10, 2010.

[28] B. D. Nye, S. Karumbaiah, S. T. Tokel, M. G. Core, G. Stratou, D. Auerbach, and K. Georgila. Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 352–366. Springer, 2018.

[29] J. Ocumpaugh. Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers*

College, Columbia University and Ateneo Laboratory for the Learning Sciences, 60, 2015.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[31] T. Sakai, M. C. du Plessis, G. Niu, and M. Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2998–3006. JMLR. org, 2017.

[32] A. Tychsen and A. Canossa. Defining personas in games using metrics. In *Proceedings of the 2008 Conference on Future Play*, 2008.

[33] S. C. Weissgerber, M.-A. Reinhard, and S. Schindler. Study harder? the relationship of achievement goals to attitudes and self-reported use of desirable difficulties in self-regulated learning. *Journal of Psychological and Educational Research*, 24(1):42, 2016.

[34] F. Wiltgren. 8 archetypes for break-testing your game, 2015.

[35] B. M. Winn. The design, play, and experience framework. In *Handbook of research on effective electronic gaming in education*, pages 1010–1024. IGI Global, 2009.

[36] N. Yee. The gamer motivation profile: What we learned from 250,000 gamers. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*, pages 2–2, 2016.