Psychological Test and Assessment Modeling, Volume 64, 2022 (4), 359-384

Identifying Aberrant Responses in Intelligent Tutoring Systems: An Application of Anomaly Detection Methods

Guher Gorgun[1] *& Okan Bulut*[2]

Abstract

Examinees' unexpected response behaviors during an assessment may lead to aberrant responses that contaminate the data quality. Since aberrant responses may jeopardize the validity of inferences made based on assessment results, they should be handled for modeling students' learning and progress more accurately. Although the detection of aberrant responses is widely studied in non-interactive low-stakes assessments, exploring aberrant responses in interactive assessment environments such as intelligent tutoring systems (ITS) is a relatively new venue. Furthermore, current aberrant response detection methods are not feasible for the ITS context due to the extreme sparsity of response data. In this study, we employed six unsupervised anomaly detection methods (Gaussian Mixture Model, Bayesian Gaussian Mixture Model, Isolation Forest, Mahalanobis Distance, Local Outlier Factor, and Elliptic Envelope) for identifying aberrant responses in an ITS environment. We compared the results of these methods with each other and explored their association with students' affective states. We found that the anomaly detection methods flagged similar responses as aberrant although Local Outlier Factor yielded very different results. Mahalanobis Distance appeared to be a conservative approach in detecting aberrant responses, whereas the Isolation Forest and Gaussian Mixture methods emerged as more liberal. Overall, the unsupervised anomaly detection methods provide a viable option for identifying aberrant responses in ITS. We recommend that researchers and practitioners consider using multiple anomaly detection methods to identify aberrant responses more accurately.

Keywords: aberrant responding, unsupervised anomaly detection, intelligent tutoring system, response time, hint use

[1] Measurement, Evaluation, and Data Science, University of Alberta

[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta *Correspondence concerning this article should be addressed to:* Guher Gorgun, Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, 6-110 Education Centre North, 11210 87 Ave, NW, Edmonton, AB, T6G 2G5, Canada, email: gorgun@ualberta.ca.

Intelligent tutoring systems (ITS) have emerged as a viable option for assessing students' learning gains while providing one-on-one tutoring to students outside of the classroom (Feng et al., 2009). ITS are interactive learning environments that aim to provide personalized instruction, assessment, and feedback to each learner. By combining tutoring assistance with assessment, researchers and practitioners can blend instruction and assessment to better support student learning. In addition to personalizing instruction and assessment, ITS also gather rich information about student behaviors such as speed, attempts to answer questions, and help-seeking, which could be stored and used for modeling complex learning behaviors. However, these systems fall under the umbrella term of low-stakes assessments. One distinctive attribute of low-stakes assessments is the lack of personal consequences or value for students. That is, although the results of low-stakes assessments may be useful for informing school accountability, funding decisions for educational institutions, and performance comparisons at the national or international levels, they are not used for higher-stakes decisions about students, such as grading students' performances, granting awards to students, or making graduation decisions (Finn, 2015). Therefore, students do not observe any direct consequences associated with low-stakes assessments and may regard them of little to no value (Lindner et al., 2019; Wise & Kong, 2005).

The tension between measuring students' knowledge and the absence of personal consequences in low-stakes assessments may lead to aberrant response behaviors due to the lack of effortful response behavior. In low-stakes assessment contexts, students with aberrant response behavior fail to show their true ability levels because they do not genuinely attempt the items by putting sufficient effort. An indication of the lack of sufficient effort is spending either an unrealistically short or long response times on each item. Researchers labeled these aberrant response behaviors with different names, such as *rapid guessing* (i.e., spending a very short response time on an item) (e.g., Rios et al., 2017; Soland & Kuhfeld, 2019; Wise, 2017); disengaged responding (e.g., Gorgun & Bulut, 2021; Wise & DeMars, 2005; Yildirim-Erbasli & Bulut, 2021); or insufficient effort responding (e.g., Liu et al., 2020; Ulitzsch et al., 2022a). Researchers have also identified other forms of aberrant responding peculiar to ITS environments. This type of response behavior is referred to as either *gaming behavior* or *off-task behavior* (e.g., Baker et al., 2004; Baker et al., 2008; Walonoski & Heffernan, 2006) in the literature. Specifically, students may exploit hints in a very short response time while attempting to answer the items correctly by abusing the system properties. In this study, we assemble all these deviant responses as *aberrant response behavior*. Accordingly, we define aberrant responses as unexpected or deviant responses from the expected response behavior or patterns (Kim et al., 2016).

In the following sections, we discuss the types and indicators of aberrant response behaviors in low-stakes assessments and methods used to identify aberrant responses. Next, we propose to harness anomaly detection methods based on unsupervised machine learning to identify aberrant responses in ITS. We argue that the convergence between multiple anomaly detection methods can help us understand the prevalence of aberrant responding in ITS and the rate of convergence across these proposed models. We evaluate the unsupervised anomaly detection methods with a publicly

available ITS dataset, ASSISTment. Our study provides preliminary evidence on the utility of unsupervised anomaly detection methods for identifying aberrant responses in ITS.

Literature Review

Aberrant Responses in Low-stakes Assessments

Aberrant responses are deviations from expected observations and are typically linked to other response mechanisms, such cheating, non-effortful test-taking, and careless responding (Hawkins, 1980; Ulitzsch et al., 2022a). Unlike responses based on genuine attempts with sufficient test-taking effort, aberrant responses may not reflect what students know or can do (Swerdzewski et al., 2011; Wise & Kong, 2005). Therefore, inferences or conclusions made based on low-stakes assessments involving aberrant responses might be unwarranted, especially if students attempt to answer the items without putting enough effort (Haladyna & Downing, 2004). In addition to unjustified inferences, absence of effortful responding may also bias statistical estimates, predictive models, classification outcomes, psychometric properties, and individual scores (e.g., DeSimone et al., 2018; Rios & Soland, 2021; Schmitt & Stuits, 1985; Wise et al., 2009). For example, Rios et al. (2017) found that scores obtained from low-stakes assessments were typically underestimated in the presence of non-effortful or disengaged responses. Similarly, students' test scores were largely distorted when they attempted the items in a disengaged mode (Wise et al., 2021).

Similar results were also reported in the context of ITS. In a recent study, Gorgun and Bulut (2022) found that when disengaged responses were removed from the training data during the pre-processing stage, the accuracy of both Bayesian and deep knowledge tracing models could significantly improve when predicting students' performances within an ITS environment. Furthermore, aberrant responses in ITS are often associated with poorer learning (Baker, 2007; Pardos et al., 2014). Due to having numerous aberrant responses, some students may even appear as if they are unlearning the content (Schultz & Arroyo, 2014). These findings corroborate the importance of identifying and handling aberrant responses to enhance not only data quality, but also the performance of predictive or psychometric models based on ITS data. Previous research indicated that aberrant responses could be handled through filtering or other correction methods (Aggarwal, 2016; Blázquez-García et al., 2021; Wise & DeMars, 2005). However, the successful detection of aberrant responses requires identifying what constitutes aberrant response behavior before performing any statistical analysis. That is, it is essential to differentiate expected and aberrant response patterns in the context of the assessment tool. In the next section, we discuss the possible manifestations of aberrant response behavior within the ITS context.

Identifying Aberrant Responses in ITS

Aberrant responses have been widely studied in low-stakes computerized assessments, including large-scale international assessments (e.g., Programme for International Student Assessment, International Reading Literacy Study, Trends in International Mathematics and Science Study; Eklöf, 2007), statewide assessments (e.g., National Assessment on Educational Progress; Braun et al., 2011; Swerdzewski et al., 2011), and computerized formative assessments used for universal screening and progress monitoring in K-12 (e.g., Agnew et al., 2021; Gorgun & Bulut, 2021; Yildirim-Erbasli & Bulut, 2021). In those studies, researchers differentiate two types of response behavior: solution behavior when the student engages with the item in an effortful fashion and non-effortful response behavior when the student does not expend maximal effort to answer the item (e.g., Pastor et al., 2019; Wise & DeMars, 2010; Wise & Gao, 2017). Rapid guessing (i.e., spending an unrealistically short response time) is the most widely studied form of non-effortful response behavior within the context of low-stakes non-interactive assessments (Deribo et al., 2021; Kroehne et al., 2020; Wise, 2017). Researchers also argued that in timed low-stakes non-interactive assessments, some students may stop investing effort and spend extremely long response times on items, leading to idle response behavior (Gorgun & Bulut, 2021). In most of these studies, response times have been used as a proxy for identifying aberrant responses. That is, non-effortfulness is operationalized through item response times.

In addition to low-stakes computerized assessments, there has been a growing interest in identifying aberrant responses in the ITS environment. However, these studies have exclusively focused on gaming behavior as a specific form of aberrant response behavior. Given the rich information that ITS can store during students' interactions within the tutoring system, researchers can use a wide variety of proxies to model different types of aberrant responses. This includes response times, the number and frequency of hint requests, the time difference between each action, the number of attempts, item difficulty, the number of errors, very fast (or slow) responses, students' first action (i.e., a hint request or an attempt), response accuracy, and students' affective states.

To date, four types of detectors have been proposed for capturing gaming behavior in ITS. Knowledge-engineered gaming detectors utilize expert knowledge to identify indicators of gaming behavior (Paquette et al., 2014). These detectors mostly focused on hint abuse and systematic guessing to model gaming behavior in ITS. In one such study, Walonoski and Heffernan (2006) used expert coded data for gaming detection. Although their anomaly detector was successful in identifying the non-gaming behavior (98% accuracy), its performance in detecting gaming behavior was quite low (19% accuracy). This may partly be explained using an imbalanced data set in terms of the number of gaming instances. Other researchers also focused on using expert coders for identifying different types of aberrant responses in the ITS environment. For example, Johns and Woolf (2006) defined three types of response behavior: motivated

response, unmotivated guess, and unmotivated hint. They employed a dynamic mixture model based on item response theory (IRT) to estimate the students' proficiency while adapting to students' motivation levels. They found that taking student motivation into account improved the model performance. Similarly, Beal et al. (2006) reported that an individual's motivation level could predict their current performance in the ITS environment better than their prior performance. Beck (2005) tried to jointly model the probability of correct response with disengagement. Although the model based on response times correctly predicted the probability of a correct response, it failed to identify disengagement.

The second type of gaming detector used machine learning to capture students' gaming behavior. Using a machine learning approach, Baker et al. (2008) developed an anomaly detector based on a latent response model for identifying when students attempt to game the system. Their anomaly detector was successful for detecting only the harmful type of gaming behavior associated with poor learning outcomes. Baker et al. (2010) also analyzed the relationship between gaming behavior and affective states of students and found that boredom and confusion were associated with poorer learning and aberrant behaviors such as system gaming. Combining the first two detectors, researchers also proposed a hybrid approach for detecting gaming behavior. Paquette et al. (2015) combined expert-coded models with machine learning and achieved comparable results to knowledge-engineered detectors. They also evaluated the model generalizability of these three approaches (i.e., knowledge-engineered, machine learning, and hybrid) across two ITS data sets. Interestingly, they found that the knowledge-engineered model achieved better generalizability to new data sets (Paquette et al., 2015). However, it is important to highlight that developing such a knowledge-engineered model may be costly (Huang et al., 2022). Finally, an extension of a knowledge-engineered approach is proposed based on IRT. In addition to expert engineered features, Huang et al. (2022) included contextual features to an IRT-based gaming detector. Their approach outperformed the previous detectors and had higher generalizability across three different data sets.

Unsupervised Machine Learning Models for Detecting Aberrant Responses

A major challenge concerning the aberrant response identification is the lack of ground truth. This may prevent researchers to evaluate the false-positive or false-negative rates, as well as to train supervised machine learning models. To address this issue, researchers often label the data with respect to aberrant responses by employing field observations or human coders (e.g., Baker et al., 2004; Walonoski & Heffernan, 2006). Several studies (e.g., Baker et al., 2004) created the labelled data through twenty-second long in-class observations of students interacting within ITS. The observers categorized students' behaviors while interacting with ITS as either on-task, off-task, inactivity, or gaming the system (for details regarding the coding process and scheme, see Baker et al., 2004). They found that 50% of students engaged in off-task

behavior and 24% of students gamed the system at least once. They also found that students who gamed the system had poorer academic performance than those who did not attempt to game the system, which is a consistent finding observed in other studies focusing on both interactive and non-interactive low-stakes assessments (e.g., Huang et al., 2022; Lindner et al., 2019).

Nonetheless, there are several possible limitations of the manual labeling process. First, when labeling aberrant responses manually, ensuring consistency within and between human coders might be very challenging. Second, creating a clear coding scheme that reflects the behavioral mechanism underlying the aberrant responses could also be difficult in practice. Third, with large-scale data sets involving thousands of learners' responses and actions in ITS, it is practically impossible to manually label all the observations with the behavior of interest (e.g., gaming the system). Finally, if the labeling procedure was to be completed through real-time observations (i.e., while students complete the items or tasks presented on ITS), some students might feel uncomfortable and thus fail to demonstrate their genuine behaviors.

To address some of the limitations mentioned above, researchers often choose to operationalize aberrant responses through proxy variables such as response time use and hint requests. However, the use of proxy variables may also constrain the types of aberrant responses to be detected, since the researcher must explicitly define what an aberrant response looks like based on the selected proxy variable (see knowledge-engineered models, e.g., Huang et al., 2022; Paquette et al., 2014). Furthermore, the current detection methods established with non-interactive assessments might not work well with interactive ITS data sets. Most of these methods require researchers or practitioners to reformat the data presentation (e.g., from long format to wide format), creating a highly sparse data set due to having many items with varying number of students taking these items (Minn, 2022). Therefore, new approaches are needed to flag students' interactions to properly identify aberrant responses in ITS.

A better approach would be to combine empirical findings (i.e., findings of previous studies on aberrant response behavior), expert knowledge (i.e., interactions within the ITS that expert coders identified as aberrant), and theoretical background (e.g., Wigfield & Eccles, 2000, Expectancy-Value Theory) with exploratory data mining techniques (i.e., unsupervised machine learning models; Romero and Ventura, 2020). This approach would allow us to identify aberrant responses in ITS when the ground truth (i.e., correct labels of aberrant responses) is unavailable or unknown. Furthermore, this theory-supported data-driven approach may help us discover new patterns of aberrant responding in the data set. Previous studies have exclusively focused on gaming behavior in ITS, but anomaly detection procedures utilizing data mining techniques can help us discover new patterns of aberrant responding.

Present Study

In this study, our primary goal is to test the feasibility and utility of various anomaly detection methods for identifying aberrant responses in interactive learning environments (e.g., ITS). Furthermore, we aim to provide further empirical evidence concerning aberrant responses and prevalence rates in a publicly available ITS data set. In addition to detecting aberrant responses, we also gather validity evidence for different anomaly detectors by two means: We analyzed the feature distribution for aberrant and normal responses to investigate whether feature distributions for both groups are congruent with the theorized expectations and we correlated students' affective states (i.e., boredom, confusion, engaged concentration, frustration) with the anomaly detection results (i.e., whether a response is identified as aberrant or not) to provide convergent validity evidence. Baker et al. (2010) found that while boredom was consistently associated with poorer learning and gaming behavior, engaged concentration is negatively related to such behavior. Thus, we anticipate finding similar trends with the aberrant responses identified through anomaly detection methods. Following Baker et al. (2010), we hypothesize that there is a positive correlation between the affective state of boredom and aberrant responses, and that there are negative correlations between affective states of frustration, engaged concentration, and confusion and aberrant responses.

The research questions of the current study are as follows:

1. What is the feature distribution for each anomaly detection method?
2. To what extent the aberrant responses identified by the anomaly detection methods overlap with one another?
3. What are the correlations between affective states (i.e., boredom, confusion, engaged concentration, frustration) and aberrant responses across different anomaly detection methods?

Method

Data

To compare the performances of different anomaly detection algorithms, we used the 2012-2013 ASSISTment school data set with affect indicators (Pardos et al., 2013)[3]. The data were collected from middle and high school students in New England, USA (Pardos et al., 2014; Pedro et al., 2013). The data set contains 36 variables including students' interactions within ITS and their predicted affective states such as boredom, confusion, frustration, and engaged concentration. The variables about students' interactions within the ITS environment involved problem start and end times, how

[3] The ASSISTment data set is available at https://sites.google.com/site/assistmentsdata/datasets/2012-13-school-data-with-affect.

many hints a student requested when answering the items and whether the student requested all the hints, response accuracy, the number of actions and actions sequences, the response time to the first action, and whether the student requested a hint or attempted the item first. There were also some variables about the problems in the ITS data set such as problem ID, whether it is a scaffolding or main question, skill ID and skill name, problem type (e.g., constructed response or multiple choice), and item position.

Data Pre-Processing

Before applying anomaly detection methods, we implemented several steps for preprocessing the data. First, we removed the duplicate rows and rows with either negative response times for the first action or missing hint information. We also removed constructed-response items (i.e., open-response problem types) because they were scored as correct regardless of the students' original responses. We retained the main problems, first actions being either an attempt or hint request, students with at least 16 items, and students' first attempt to each problem. The response times were converted from milliseconds to seconds. Furthermore, we engineered some additional features before applying anomaly detection methods. First, we found the total response time—the total response time as the gap between the problem end time and start time. We also standardized the number of hints the students requested because the problems have a different number of total hints available. In addition to standardizing hint counts, we estimated the hint requests per seconds to scale the hint (ab)use. Finally, some problems in the ITS data set are partially scored. The partial scores were assigned when the student's first attempt was incorrect, and the student requested help. Following Wang et al. (2010), we recoded responses smaller than 1 (i.e., partially scored responses) as 0 (i.e., incorrect). The final data set included 33,825 unique students with features including student ID, problem ID, hint count per second, whether the first action was an attempt or a hint request, response accuracy, attempt count, the total response time, the response time to first action, the standardized total number of hints requested, and students' affective states (i.e., predicted frustration, boredom, engaged concentration, and confusion).

Anomaly Detection Methods

To answer our research questions, we employed density-based, point-based, parametric, and distance-based anomaly detection methods. Below, we briefly describe the six unsupervised anomaly detection methods employed in this study. For each method, aberrant responses are generally defined as extreme deviations (i.e., anomaly) from a normative response behavior. Yet, nuances of each method will be described below.

Gaussian Mixture Model. Gaussian Mixture Model (GMM; Bouman et al., 1997; McLachlan & Basford, 1988) is a probabilistic density-based model that posits different data generation processes for normal and deviant (i.e., aberrant) responses. It assumes a mixture of several Gaussian distributions with unknown parameters. Each distribution may have a different shape, size, or density. The researcher must define the number of *k* Gaussian distributions. That is, the researcher may need to tune the model by finding the optimal *k* clusters through selecting a model that minimizes the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC). Furthermore, the researcher needs to define a density threshold. Therefore, this anomaly detection method can be used for situations where empirical evidence concerning the contamination rate is already known. Furthermore, this method can be employed to evaluate the convergence across different anomaly detection models to gather further validity evidence for the observations flagged as aberrant.

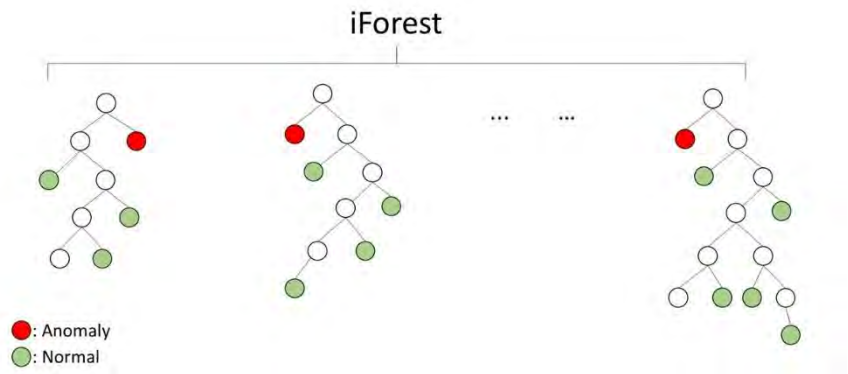
Using the expectation-maximization (EM) algorithm, GMM starts by finding the cluster parameters randomly and iteratively until convergence. For each data point, the probability of being in a given cluster is estimated to find which cluster each data point belongs. Any observation found in the low-density region is considered as a deviant observation. Given the starting values of the EM algorithm, EM may converge poorly, and thus it is advised to run the EM algorithm multiple times to find the optimal solution. That is, the researcher needs to set *the number of iterations* to run the model several times to find the best solutions. Furthermore, the researcher may also inspect whether the model has converged and the number of iterations it takes to converge. A final hyperparameter that the researcher should define is covariance type (i.e., the shape, size, or orientation that each cluster can take). In this study, we used *full* covariance type, allowing each cluster to take any shape, size, or orientation. The algorithm assigns an anomalous status to the observations that have a low probability of being in a cluster. Specifically, the density threshold determines the cut-off value for observations classified as anomalous and observations below that percentile are flagged as aberrant.

Bayesian Gaussian Mixture Model. Bayesian Gaussian Mixture Model (BGMM; Bishop & Nasrabadi, 2006) is another probabilistic density-based Gaussian method, which has some advantages over GMM for anomaly detection. Unlike manually determining the optimal number of clusters in GMM, BGMM can determine the optimal number of clusters through giving zero weights to the unnecessary clusters. The researcher needs to set the hyperparameter *the number of components* to a value that is expected to be higher than the anticipated number of optimal clusters. Similar to GMM, the researcher needs to run the algorithm several times to find the best solution. However, one significant drawback of BGMM over GMM is that it is computationally more intensive, hence it takes a longer time to converge. For non-ellipsoidal clusters, BGMM may fail to identify the underlying number of clusters properly and may perform bad. For such non-ellipsoidal clusters, several other algorithms (e.g., Isolation Forest or Local Outlier Factor) may perform better than GMM and BGMM. Similar to GMM, observations are flagged as aberrant if they are below the researcher defined density threshold.

Isolation Forest. Isolation Forest (iForest; Liu et al., 2008) is robust point-based anomaly detection algorithm, which has been shown to work well with high dimensional data. iForest is a tree-based ensemble method and identifies anomalies with the assumption that anomalies are less frequent and distinct from normal observations. Typically, during anomaly detection, norms are identified first then the deviations from the normal observations are flagged as anomalous. However, iForest tries to isolate deviations from normal observations first. Liu et al. (2008) observed that normal data points require more partitions compared with anomalous data points. Hence, the algorithm starts by selecting a feature and then randomly picking a value between the maximum and minimum values of the feature to split the tree (see Figure 1). In Figure 1, the nodes represent data points. The algorithm repeatedly partitions the random trees until all instances are isolated. The shorter paths in the tree structure are conceptualized to be anomalies because the algorithm finds it easier to distinguish them from the other observations (Liu et al., 2008). Thus, data points with shorter paths (e.g., red nodes in Figure 1) are classified as anomalous data points.

Figure 1

The iForest Algorithm



Mathematically, iForest can be expressed as

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (1)$$

where *h(x)* is the path length of observation *x*, *c(n)* is the average path length of an unsuccessful search and *n* is the number of external nodes. Given the formula, a score of an anomaly is assigned to each observation, where scores closer to 1 indicate anomalies. Scores smaller than 0.5 indicates normal observations. When training iForest, the researcher may set a contamination rate, however, in this study we set the contamination rate to *auto* and allowed the algorithm to freely discover the anomalous points in the data set.

Mahalanobis Distance. Mahalanobis Distance (MD; Mahalanobis, 1936) is a distance-based anomaly detection method, which has been widely used as a multivariate outlier detection method (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012; Ulitzsch et al., 2022b). This method creates a distance metric based on each variable in the data set. MD works well with multivariate data and, using covariance matrix among variables, it calculates the distance between a data point and the center. Formally, this distance is calculated as

$$D^2 = (x-m)^T C^{-1}(x-m), \quad (2)$$

where *C* is the covariance matrix, *x* is the vector of observations and m is the vector of mean values. We determined the cut-off value for MD by taking the quantile value of 0.95 and the points beyond this cut-off value were considered as anomalies. MD shows superior performance compared with the Euclidean distance in the presence of skewed data because MD considers the covariance among the variables.

Local Outlier Factor. The Local Outlier Factor (LOF; Breunig et al., 2000) method is a widely used density-based method and calculates the local deviations relative to its surrounding points. Specifically, LOF identifies anomalies based on the density of the neighborhood. This method works well for anomaly detection for data sets with uneven distributions (Cheng et al., 2019). Similar to GMM and BGMM, a data point can be identified as anomalous if it is in a lower density region. The LOF method first finds the *k*-nearest neighbors for data points. Then, using the *k*-nearest neighbors, a local density metric named the local reachability density (LRD) is calculated. Formally defined, LRD can be computed as

$$LDR_k(x) = 1/\left(\frac{\sum_{o\epsilon N_k(x)} d_k(x,o)}{|N_k(x)|}\right), \quad (3)$$

where *x* is a data point, *o* is *k*-distance, and *N* is the minimum number of data points.

Finally, the LOF score for a data point is found by comparing the LDR with its *k*-neighbors. LOF can be given as

$$LDR_k(x) = \frac{\sum_{o \epsilon N_k(x)} \frac{LDR_k(o)}{LDR_k(x)}}{|N_k(x)|}, \quad (3)$$

where the LOF score greater than 1 indicates anomalies. This method is suitable for identifying local anomalies because the method uses the *k*-nearest neighbor of a data point to identify whether it is an anomaly or not. Therefore, it is important to select the optimal number of clusters (i.e., *k*) where the underlying anomaly pattern reflects the local outliers rather than the global ones. Furthermore, a drawback of the LOF approach is that the number of neighbors must be manually set by the researcher. Therefore, hyper-tuning the number of neighbors might be challenging without any prior information on the data set.

Elliptic Envelope. The Elliptic Envelope (EE; Rousseeuw & Driessen, 1999) method is a simple method that considers all observations simultaneously, rather than each individual feature separately. EE assumes a normal distribution with covariance among the features. Basically, EE tries to learn an optimized boundary of imaginary ellipse separating normal data points (i.e., inliners) from anomalous data points (i.e., outliers). The optimal imaginary ellipse is learned with FAST-Minimum Covariance Determinant algorithm introduced by Rousseeuw and Driessen (1999). The anomaly detection process is initialized by selecting non-overlapping subsamples of data. Then, MD is calculated for each subsample with the goal of measuring deviations between each data point and the mean of the distributions of the data. In the final step, the imaginary ellipse is determined by selecting the covariance matrix with the smallest determinant in all subsamples. The data points that are far away from the shape coordinates are considered as anomalous data points. The researcher needs to set *the contamination rate* as a hyperparameter.

Using the anomaly detection methods summarized above, we performed an anomaly detection analysis based on the features identified in the ASSISTment data set and compared the results in terms of the number of anomalous cases (i.e., aberrant responses) identified by each method. All analyses were conducted using the sklearn library (Pedregosa et al., 2011) in Python (version 3.10.4; Python Software Foundation, 2022).

Results

Aberrant Responses and Feature Distributions

With the contamination rate identified as 18% based on previous studies, GMM flagged 18% of data points as anomalous (i.e., aberrant). During the training process, we set the contamination rate, the number of Gaussian distributions, and the number of iterations. Based on the AIC and BIC values, the best number of Gaussian distributions was 6. With GMM, we can also check whether the model was converged, and

the number of iterations required for convergence. In this study, GMM converged with 43 iterations. One advantage of GMM is that the density thresholds differentiating anomalies from normal data points could be set. That is, it is possible to decrease the threshold to avoid too many false positives or too many false negatives when the contamination rate is known (Géron, 2019).

Similar to GMM, we set the contamination rate when using BGMM and 18% of the data points were flagged as aberrant. An advantage of BGMM over GMM is that instead of manually searching for the optimal number of clusters, BGMM can automatically identify the optimal number of clusters. That is, after starting with a large number of clusters at the beginning, the model automatically tunes itself to find the optimal number of clusters. Therefore, we set the number of Gaussian distributions to a high number of clusters (i.e., k = 20) and the model identified the optimal number of clusters to be 6. However, both GMM and BGMM require a density threshold (i.e., percentage of anomalies that the researcher expect in a data set or a contamination rate) to be manually set, which might be challenging when these methods are applied to new, unexplored venues. In this study, we set the contamination rate informed by previous studies (e.g., Baker et al., 2004; Huang et al., 2022).

iForest identified 17% of the data points as aberrant responses. We set the number of samples to be drawn from the data and contamination rate to *auto* in order to follow a purely data-driven approach. However, if rates of anomalies or aberrant responses are already known, it is possible to set these values congruent with those from previous empirical studies. Compared with GMM and BGMM, iForest is more robust and easier to use because hyperparameters such as the number of clusters or the contamination rate (i.e., aberrant response rate) do not have to be defined manually.

MD identified 9% of the data points as anomalous. This proportion was much smaller than the observed rate of aberrant responses in previous studies on ITS (Baker et al., 2004; Walonoski & Heffernan, 2006). To evaluate why this discrepancy occurred between MD and other anomaly detection methods (i.e., GMM and iForest), we analyzed the correlations among the features. Except for a few features, the features used in modeling MD did not strongly correlate with one another. This may partially explain why we observed a lower prevalence rate of aberrant responses with MD. It also implies that this method could be considered as a conservative method.

With LOF, 9% of the data points were flagged as aberrant, which was the same proportion of responses flagged with MD and it was significantly lower than the proportion of responses flagged by GMM and iForest. With the EE method, we flagged 18% of the data points as aberrant. While considering all features as a whole is an advantage of this anomaly detection algorithm, a significant disadvantage is the need to predetermine the contamination rate. Based on the previous studies involving the ITS data (Baker et al., 2008; Huang et al., 2022), we set the contamination rate to 18%, so it is not surprising to have 18% of the data flagged as aberrant.

Table 1

The Feature Means for Aberrant and Normal Responses across the Anomaly Detectors

Anomaly Detector	RT		Hint (B) (%)		Hint (S)		Hint (F) (%)		Correct (%)		Attempt	
	A	N	A	N	A	N	A	N	A	N	A	N
GMM	83.66	41.91	64	1	1.02	-0.23	23	1	11	81	2.31	1.14
BGMM	87.70	41.02	54	3	0.86	-0.19	13	3	6	82	2.55	1.09
iForest	110.33	37.07	61	2	0.97	-0.20	28	0	18	79	2.28	1.17
MD	70.67	47.43	78	6	1.50	-0.14	55	0	1	75	2.58	1.24
LOF	55.51	48.79	8	12	0.02	0	4	5	71	68	1.32	1.36
EE	71.62	50.25	52	3	0.85	-0.19	17	2	1	84	2.54	1.09

Note. BGMM: Bayesian Gaussian Mixture Model; EE: Elliptic Envelope, GMM: Gaussian Mixture Model; iForest: Isolation Forest; LOF: Local Outlier Factor; MD: Mahalanobis Distance; RT: Response Time; Hint (B): whether the student requested all hints; Hint (S): Standardized Hint Count; Hint (F): whether the student requested a hint or attempt first; A: Aberrant Response; and N: Normal Response.

We provided further validity evidence for the utility of these algorithms by analyzing the feature distributions for aberrant and normal responses for each of these methods. This type of validation is similar to the knowledge-engineered detector developed for ITS (see Huang et al., 2022; Paquette et al., 2014). Our findings concerning the feature distributions are reported in Table 1. All detectors but LOF performed as expected. The response time use was higher for aberrant responses. This is an expected outcome since students with aberrant responses tend to spend more time and engage in idle responding behavior (see Baker, 2007). The mean response time for GMM and BGMM were almost identical whereas we observed different mean response times for different anomaly detection methods. Nonetheless, across all the methods, except for LOF, the mean response time was longer for aberrant responses than normal responses. Furthermore, we observed that the students with aberrant responses requested more hints (see Hint (B) and Hint (F) in Table 1) except for aberrant responses identified with LOF. For students with aberrant responses, the percent correct scores were much lower, and the average attempt count was much higher than normal responses. While these results provide empirical evidence for the utility of anomaly detection methods of GMM, BGMM, iForest, MD, and EE, the feature distribution for LOF is alarming given that the most feature distributions were against the postulated behaviors of disengaged students .

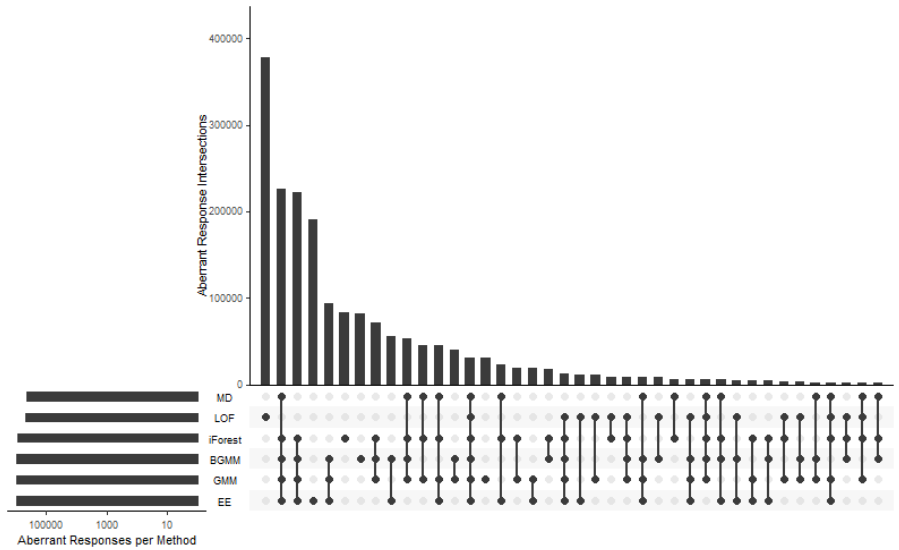
Overlap between the Anomaly Detectors

Comparing the overlap among the data points identified as anomalous (i.e., aberrant) with each algorithm can help us better understand the convergence rates between different anomaly detectors and provide stronger evidence regarding which data points were truly aberrant. The UpSet plot in Figure 2 demonstrates the intersections (i.e., overlap) among the anomaly detectors in terms of the number of aberrant responses detected in the ITS data set. Among the six detectors, LOF flagged the largest number of unique data points as aberrant responses (> 380, 000), followed by EE (> 190, 000) and iForest (> 85, 000).

The aberrant responses detected by the LOF method were almost entirely different, compared with those detected by the other anomaly detection methods. Despite the unique data points, the UpSet plot shows that aberrant responses detected with MD, iForest, GMM, and BGMM overlapped greatly with one another. There are also a number of data points identified as aberrant responses by more than four detectors. However, there is no clear pattern in terms of the overlapping detectors. We also observed that aberrant responses identified by MD were almost a proper subset of GMM and iForest. This trend was less evident with BGMM and EE.

Figure 2

The overlap between Six Anomaly Detectors



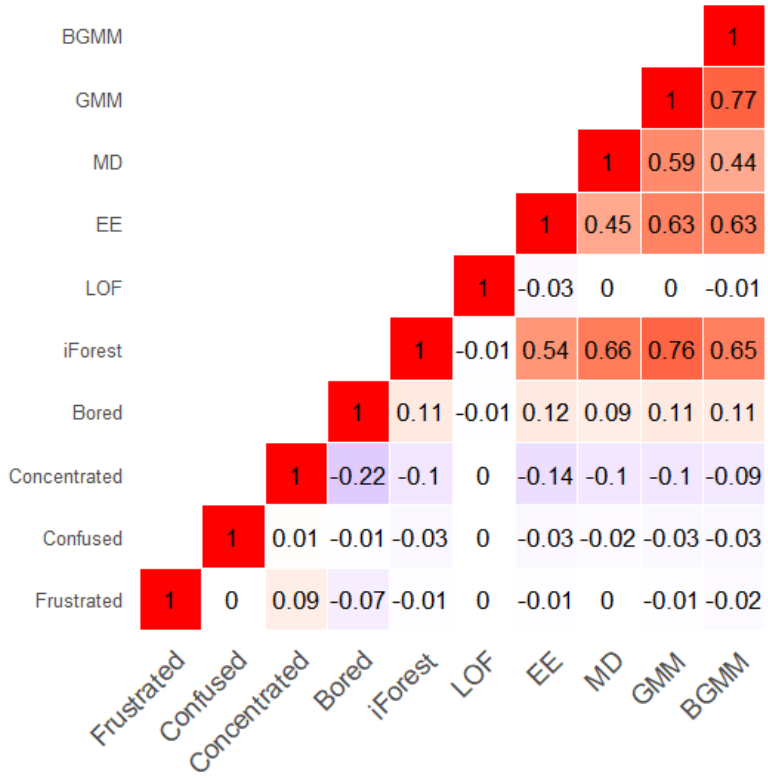
Note. BGMM: Bayesian Gaussian Mixture Model; EE: Elliptic Envelope, GMM: Gaussian Mixture Model; iForest: Isolation Forest; LOF: Local Outlier Factor; MD: Mahalanobis Distance. Aberrant responses per method were rescaled with log10.

Convergence between Affective States and Aberrant Responses

We evaluated the convergence between the affective states and aberrant responses to provide convergent validity evidence. First, we analyzed the correlations among anomaly detectors (see Figure 3). We found strong correlations among iForest, GMM, BGMM, EE, and MD (i.e., ranging from r = .44 to r = .76). The results of all anomaly detection methods, except for LOF, seemed to be correlated with each other. This was an interesting finding that requires further investigation to decipher the underlying reasons for this discrepancy.

Figure 3

Correlations among Affective States and Anomaly Detection Methods



Note. BGMM: Bayesian Gaussian Mixture Model; EE: Elliptic Envelope, GMM: Gaussian Mixture Model; iForest: Isolation Forest; LOF: Local Outlier Factor; MD: Mahalanobis Distance.

In line with previous studies focusing on gaming behavior within the ITS context, we found a significant negative correlation between aberrant responses and engaged concentration. Although the strength of the correlation was small (i.e., $r \approx .10$), we observed the same trend for EE, iForest, GMM, BGMM, and MD. Furthermore, we observed positive significant correlations between boredom and aberrant responses across the same anomaly detection methods. However, the results of these anomaly detection methods were not correlated with confusion and frustration. Overall, these findings are in accordance with findings reported by Baker et al. (2008). Interestingly, we did not expect to observe non-significant correlations between LOF, engaged concentration, and boredom. Since LOF did not correlate well with other anomaly detection methods and affective states, it is possible that either LOF identifies a completely different set of anomalies, or it is heavily influenced by the distributional properties of the data set or features. We tuned LOF with different number of k neighbors (i.e., $k = [2:6]$) and analyzed the correlations between LOF, affective states, and other anomaly detectors, yet we obtained very similar results across LOF with different number of k neighbors. Note that the way LOF identifies the local deviations, and its underlying mechanism may be completely different from other unsupervised anomaly detection algorithms used in this study. This may also suggest that aberrant response patterns are global rather than local, hence methods focusing on local anomalies (e.g., LOF) may not be apt for identifying aberrant responses in ITS.

Discussion

In this study, we compared the performances of six anomaly detection methods for identifying aberrant responses in an interactive cognitive assessment environment (i.e., ITS). The absence of ground truth about aberrant responses complicates the identification of aberrant responses in the ITS data. Similarly, non-interactive cognitive assessments also lacked the ground truth about aberrant responses. Thus, the approaches used for aberrant response detection in both interactive and non-interactive assessment environments typically are unsupervised methods. However, detecting aberrant responses in non-interactive environments is less challenging because data sets from non-interactive assessments have fewer number of items with relatively fewer number of missing responses. Interactive assessments, on the other hand, are likely to have hundreds of items, some of which are administered multiple times whereas others are administered only a few times.

This feature of interactive assessments complicates the aberrant response detection in ITS with methods established for non-interactive cognitive assessments and renders most of the psychometric approaches used for aberrant response detection useless. The performance of aberrant response detection methods (e.g., Wise & Ma, 2012, threshold-based methods) used with non-interactive assessments may be evaluated through statistical indices (e.g., improvement in fit statistics after aberrant responses are removed Rios & Soland, 2021; Ulitzsch et al., 2020; Wise, 2017); however, these approaches are not applicable to interactive learning environments. Therefore, the

anomaly detection methods illustrated in this study are more promising for flagging aberrant responses in ITS. To the best of our knowledge this is the first study comparing six different anomaly detection methods and providing validity evidence through analyzing feature distributions and convergence between affective states and anomalies.

An interesting finding is that the results from the anomaly detection methods used in this study were very similar, except for the results of LOF. LOF identified a different set of responses as aberrant and performed very differently compared with the other anomaly detection algorithms. Furthermore, LOF was not adequately correlated with affective states present in the ITS data set. Affective states worked as our convergent validity criteria to evaluate the performances of anomaly detection methods in this study. We expected to observe significant associations between affective states and aberrant responses. The performance of LOF was contradictory to what we expected to observe; however, it highlighted the importance of selecting proper anomaly detection algorithms to capture the underlying mechanism of aberrant responses. Thus, researchers and practitioners should corroborate their choice of aberrant response detector by replicating the results through other methods and means. Researchers may set the contamination rate informed by other anomaly detection methods to evaluate the congruence between the selected methods. This would allow researchers to evaluate whether the same instances are flagged as aberrant when the same contamination rate is used across the anomaly detection methods proposed in this study.

In this study, we found that iForest, EE, MD, GMM, and BGMM performed very similarly. We further observed that MD was almost a proper subset of these methods, and thus it identified a smaller set of responses as aberrant. These findings suggest that, when researchers want to avoid false positives in detecting aberrant responses, they can choose MD during the data cleaning or pre-processing stages. However, if the researcher aims to remove as many aberrant responses as possible (e.g., Soland et al., 2021), then the other anomaly detection algorithms (i.e., iForest, EE, GMM and BGMM) may be better options. For example, when the goal is to make inferences about one's performance, researchers or practitioners may choose a more conservative method (i.e., MD); however, when the goal is to develop models about students' learning and progress at the aggregate level (Gorgun & Bulut, 2022), they may opt for more liberal methods (i.e., iForest). Depending on the researchers' and practitioners' needs, the overlapping aberrant data points may be used to achieve accurate yet more certain aberrant response flagging. Nonetheless, some situations may call for a more liberal method where a data point is considered aberrant whenever it is flagged as anomalous with one of the methods used in this study. In this respect, researchers or practitioners may benefit from highlighting the use and inferences that they want to make by the use of the assessment. Considering the use and interpretations of assessments may help researchers or practitioners pick the appropriate anomaly detection methods.

This study, to the best of our knowledge, is the first study comparing the performances of various anomaly detection algorithms based on unsupervised machine learning for identifying aberrant responses in ITS. We cannot pinpoint the best anomaly detection

method since the ITS data set utilized in this study is very contextual and the ground truth about aberrant responses is also absent. However, we argue that the convergence between different anomaly detection methods and the observed correlations between affective states and aberrant responses show that the anomaly detection algorithms of iForest, EE, MD, GMM, and BGMM could be more promising for detecting aberrant responses in ITS. We recommend that researchers combine multiple anomaly detection algorithms with other possible proxies in the data set (e.g., response time use, affective states) to flag aberrant responses more accurately and consistently. The lack of ground truth is a significant factor complicating the aberrant response detection. Therefore, we also recommend conducting some descriptive analysis of feature distributions for both normal and aberrant data points to support aberrant response flagging. This would allow researchers and practitioners to support aberrant response flagging with theoretical expectations. For instance, we typically expect to observe that aberrant responses to have higher hint requests than normal responses, reflecting hint-abuse behavior. Corroborating the aberrant response identification process with descriptive statistics can help researchers and practitioners validate their findings.

Limitations and Future Direction

Our study has several limitations. First, the absence of ground truth (i.e., which responses are truly aberrant) restricts us to employ only unsupervised anomaly detection methods. This situation also limited the model evaluation methods that we could use for understanding the false-positive and false-negative rates in the data set. Future studies could utilize ITS data sets with labelled aberrant responses to further evaluate the performances of the anomaly detection methods. Second, ITS data sets with aberrant responses are typically imbalanced, since the number of aberrant responses are much smaller than the number of non-aberrant responses. In this study, we did not consider the impact of the potential class imbalance problem on anomaly detection methods. Although some algorithms utilized in the study can handle the class imbalance problem effectively (e.g., iForest; Vikram et al., 2020), the others might have been influenced by the presence of imbalanced classes in the ITS data set. Future research is necessary to investigate the performances of anomaly detection methods when the class imbalance problem is addressed through sampling approaches such as undersampling, oversampling, and SMOTE. Third, we used the students' first attempt to each question when identifying aberrant responses. This may have restricted the types of aberrant responses identified in the ITS data set. Future research may extend the findings of this study by considering multiple attempts made for each item in the ITS. Finally, we did not consider item or person characteristics when identifying the aberrant responses with the unsupervised anomaly detection methods. Yet, person (e.g., ability) and item characteristics (i.e., difficulty) may have influenced the performances of the methods employed in this study. Future research can investigate the potential effects of such characteristics to better understand the utility of the proposed unsupervised anomaly detection methods for identifying aberrant responses.

References

Aggarwal, C. C. (2016). *Outlier analysis*. Springer.

Agnew, S., Kerr, J., & Watt, R. (2021). The effect on student behaviour and achievement of removing incentives to complete online formative assessments. *Australasian Journal of Educational Technology, 37* (4), 173–185. https://doi.org/10.14742/ajet.6203

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students" game the system". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 383–390. https://doi.org/10.1145/985692.985741

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18* (3), 287–314. https://doi.org/10.1007/s11257-007-9045-6

Baker, R. S. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* 1059–1068. https://doi.org/10.1145/1240624.1240785

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). 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. https://doi.org/10.1016/j.ijhcs.2009.12.003

Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. *AAAI*, 151–156.

Beck, E., J. (2005). Engagement tracing: Using response times to model student disengagement. In C-K Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology* (pp. 88-95). IOS Press.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR), 54* (3), 1–33. https://doi.org/10.1145/3444690

Bouman, C. A., Shapiro, M., Cook, G., Atkins, C. B., & Cheng, H. (1997). Cluster: An unsupervised algorithm for modeling gaussian mixtures. Purdue University. http://dynamo.ecn.purdue.edu/~bouman/software/cluster

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113* (11), 2309–2344. https://doi.org/10.1177/016146811111301101

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104. https://doi.org/10.1145/342009.335388

Cheng, Z., Zou, C., & Dong, J. (2019). Outlier detection using isolation forest and local outlier factor. *Proceedings of the Conference on Research in Adaptive and Convergent Systems,* 161–168. https://doi.org/10.1145/3338840.3355641

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement, 58*(2), 281–303. https://doi.org/10.1111/jedm.12290

DeSimone, J. A., DeSimone, A. J., Harms, P., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology, 67*(2), 309–338. https://doi.org/10.1111/apps.12117

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311–326. https://doi.org/10.1080/15305050701438074

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, 2015(2), 1–17. https://doi.org/10.1002/ets2.12067

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. Educational and Psychological Measurement, *81*(5), 847–871. https://doi.org/10.1177/0013164421991211

Gorgun, G., & Bulut, O. (2022). Considering disengaged responses in Bayesian and deep knowledge tracing. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners' and doctoral consortium* (pp. 591-594). Lecture Notes in Computer Science, vol 13356. Springer, Cham. https://doi.org/10.1007/978-3-031-11647-6_122

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hawkins, D. (1980). *Identification of outliers*. Springer.

Huang, Y., Dang, S., Richey, J. E., Asher, M., Lobczowski, N. G., Chine, D., McLaughlin, E. A., Harackiewicz, J. M., Aleven, V., & Koedinger, K. (2022). Item response theory-based gaming detection. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th international conference on educational data mining* (pp. 251–262). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853093

Johns, J., & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. *AAAI,* 163–168.

Kim, D., Woo, A., & Dickison, P. (2016). Identifying and investigating aberrant responses using psychometrics-based and machine learning-based approaches. In *Handbook of quantitative methods for detecting cheating on tests* (pp. 70–97). Routledge.

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147–177. https://doi.org/ 10.25656/01:23630

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology, 10*, 1533. https://doi.org/10.3389/fpsyg.2019.01533

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *8th IEEE International Conference on Data Mining,* 413–422.

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement, 80*(4), 775–807. https://doi.org/10.1177/0013164419895068

Mahalanobis, P. C. (1936). On the generalized distance in statistics.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). M. Dekker New York.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437. https://doi.org/ 10.1037/a0028085

Minn, S. (2022). AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence*, *3*, 100050. https://doi.org/10.1016/j.caeai.2022.100050

Paquette, L., Baker, R.S., de Carvalho, A., Ocumpaugh, J. (2015). Cross-System Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System. In Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (Eds) *User modeling, adaptation and personalization. UMAP 2015. Lecture notes in computer science* (vol 9146). Springer, Cham. https://doi.org/10.1007/978-3-319-20267-9_15

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. *CogSci.*

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 117–124.

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics, 1*(1), 107–128.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment, 24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of Educational Data Mining Conference.*

Python Software Foundation. (2022). *Python language reference* (Version 3.10.4). https://www.python.org

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the effort-moderated irt model serve as a solution? *Applied Psychological Measurement, 45*(6), 391–406. https://doi.org/10.1177/01466216211013896

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355.

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schultz, S., & Arroyo, I. (2014). Tracing knowledge and engagement in parallel in an intelligent tutoring system. *Proceedings of Educational Data Mining Conference.*

Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment, 24*(4), 327–342. https://doi.org/10.1080/10627197.2019.1645592

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education, 9*(1), 1–21. https://doi.org/10.1186/s40536-021-00100-w

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022a). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology, 73*, 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology, 75*(3), 668-698. https://doi.org/10.1111/bmsp.12272

Vikram, A., et al. (2020). Anomaly detection in network traffic using unsupervised machine learning approach. *5th International Conference on Communication and Electronics Systems (ICCES)*, 476–479.

Walonoski, J. A., & Heffernan, N. T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. *International Conference on Intelligent Tutoring Systems,* 382–391.

Wang, Y., Heffernan, N. T., & Beck, J. E. (2010). Representing student performance with partial credit. *Proceedings of Educational Data Mining Conference*, 335–336.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment, 26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a cat item pool: The normative threshold method [Paper presented at Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada].

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Yildirim-Erbasli, S. N., & Bulut, O. (2021). The impact of students' test-taking effort on growth estimates in low-stakes educational assessments. *Educational Research and Evaluation, 26*(7-8), 368–386. https://doi.org/10.1080/13803611.2021.1977152

Appendix

Python Code

```
//Libraries needed
import numpy as np
from collections import Counter
//Gaussian Mixture Model
from sklearn.mixture import GaussianMixture
gmm = GaussianMixture(n_components = 6, n_init =15, rand
om_state = 42)
gmm_pred=gmm.fit_predict(df)
gmm.converged_ //check whether the model converged
gmm.n_iter_ //number of iterations it takes to converge
densities = gmm.score_samples(df)
density_threshold = np.percentile(densities, 18)
//here we set the contaminate rate to 18%
anomalies_gmm = df[densities < density_threshold]
//data points smaller than density threshold were flagge
d as anomalous
//Bayesian Gaussian Mixture Model (BGMM)
from sklearn.mixture import BayesianGaussianMixture
bgmm = BayesianGaussianMixture(n_components=20, n_init=1
5)
bgmm.fit(df)
densities = bgmm.score_samples(df)
density_threshold = np.percentile(densities, 18)
//here we set the contaminate rate to 18%
anomalies_bgmm = df[densities < density_threshold]
//data points smaller than density threshold were flagge
d as anomalous
//Isolation Forest (iForest)
from sklearn.ensemble import IsolationForest
iForest = IsolationForest(max_samples='auto', contaminat
ion= 'auto')
//here we set the hyperparameters max_samples and contam
ination
```

```
iForest_pred = iForest.fit_predict(df)
counter = collections.Counter(iForest_pred)
iForest_pred_anomaly=[1 if i == -1 else 0 for i in preds
]
//anomalies were coded with 1
//Mahalanobis Distance (MD)
distances = []
for i, val in enumerate(df):
     p1 = val
     p2 = centerpoint
     distance = (p1-p2).T.dot(covariance_pm1).dot(p1-p2)
     distances.append(distance)
distances = np.array(distances)
cutoff = chi2.ppf(0.95, df.shape[1])
//Threshold value for detecting anomalies
outlier_index = np.where(distances > cutoff ) //index of
anomalies
np.count_nonzero(outlier_index) //number of anomalies
df["mahalanobis"] = np.nan
df['mahalanobis'] = df.loc[outlier_index,'mahalanobis']
== 1
df['mahalanobis'] = [1 if i == False else 0 for i in df[
'mahalanobis']]
//anomalies were coded with 1
//Local Outlier Factor (LOF)
from sklearn.neighbors import LocalOutlierFactor
lof = LocalOutlierFactor(n_neighbors=6, contamination =
'auto', novelty = False)
//here we need to set the hyperparameter n_neighbors
lof_pred = lof.fit_predict(df)
lof_pred_anomaly=[1 if i == -1 else 0 for i in lof_pred]
//anomalies were coded with 1
//Elliptic Envelope (EE)
from sklearn.covariance import EllipticEnvelope
elliptic = EllipticEnvelope(contamination =.18)
//here we need to set the hyperparameter contamination
ee_pred=elliptic.fit_predict(df)
ee_pred_anomaly=[1 if i == -1 else 0 for i in ee_pred]
//anomalies were coded with 1
```