Article

Applied Psychological Measurement
2020, Vol. 44(7-8) 499–514

DOI: 10.1177/0146621620931198
journals.sagepub.com/home/apm
SAGE

Using Machine Learning Methods to Develop a Short Tree-Based Adaptive Classification Test: Case Study With a High-Dimensional Item Pool and Imbalanced Data

Yi Zheng[1], Hyunjung Cheon[1] and Charles M. Katz[1]

Abstract

This study explores advanced techniques in machine learning to develop a short tree-based adaptive classification test based on an existing lengthy instrument. A case study was carried out for an assessment of risk for juvenile delinquency. Two unique facts of this case are (a) the items in the original instrument measure a large number of distinctive constructs; (b) the target outcomes are of low prevalence, which renders imbalanced training data. Due to the high dimensionality of the items, traditional item response theory (IRT)-based adaptive testing approaches may not work well, whereas decision trees, which are developed in the machine learning discipline, present as a promising alternative solution for adaptive tests. A cross-validation study was carried out to compare eight tree-based adaptive test constructions with five benchmark methods using data from a sample of 3,975 subjects. The findings reveal that the best-performing tree-based adaptive tests yielded better classification accuracy than the benchmark method IRT scoring with optimal cutpoints, and yielded comparable or better classification accuracy than the best benchmark method, random forest with balanced sampling. The competitive classification accuracy of the tree-based adaptive tests also come with an over 30-fold reduction in the length of the instrument, only administering between 3 to 6 items to any individual. This study suggests that tree-based adaptive tests have an enormous potential when used to shorten instruments that measure a large variety of constructs.

Keywords

adaptive test, classification tree, machine learning

In the field of psychometrics, the need for a short and efficient measurement instrument is ubiquitous. For example, in mental health diagnosis and risk assessment, many traditional measurement instruments are lengthy and take hours to complete. The target population, however, typically sick patients, children, or youths, is often unsuitable to answer lengthy questionnaires.

[1]Arizona State University, Tempe, USA

Corresponding Author:
Yi Zheng, Mary Lou Fulton Teachers College and School of Mathematical and Statistical Sciences, Arizona State University, 1050 S. Forest Mall, Tempe AZ 85281, USA.
Email: yi.isabel.zheng@asu.edu

The total administration cost is another concern for large-scale implementation. Hence, the idea of shortening an existing lengthy measurement instrument is often highly appealing. The best-known technique for creating a short yet efficient measurement instrument is probably *computer/computerized adaptive testing* (CAT). As its name implies, this technique uses computer software to adapt a test to each individual test-taker, selecting the most suitable items in real time. Although such an intuitive concept of CAT is quite versatile, the last half a century has witnessed the dominance of CAT designs based on *item response theory* (IRT; Hambleton et al., 1991; for example, Chang, 2015; Wainer, 2000). After decades of large-scale implementation of IRT-based CAT programs,[1] the term ''computer adaptive testing'' or ''CAT'' is now understood by many psychometricians as almost exclusively referring to IRT-based CAT. IRT is one of the most pivotal inventions of contemporary measurement theory to date, and has enabled many new and powerful ways to analyze and design measurement instruments. Yet it is not always a good choice.

One critical limitation of IRT is it is a strong model that relies on the strict assumption of dimensionality. In other words, a unidimensional IRT model is only adequate for measurement instruments where all items share sufficient communality, for instance, a state anxiety scale (Marteau & Bekker, 1992) where all questions assess state anxiety, only from slightly different angles. Multidimensional IRT models can accommodate more than one dimension. Nevertheless, they also require that each dimension be well supported by a set of tightly coherent items (i.e., internal consistency). In practice, many measurement applications may fall short of such requirement. One example is educational tests that measure a large and diverse knowledge domain. Typically, there is a blueprint that lays out the various topics, and the items are written or assembled to cover all topics. This practice, designed to establish content validity of the test (American Educational Research Association [AERA] et al., 2014), often results in items that are highly diverse and do not share enough communality to be accurately modeled by IRT. Hence, the adequacy of an IRT-based CAT is questionable.

Another example, which is the case the authors are studying in this article, is risk assessment. The objective of the risk assessment in the studied case is to screen for juveniles at risk for delinquency in a Central American country. The incumbent screening process is carried out using a questionnaire consisting of 38 scales, totaling 173 items. These 38 scales collect information on distinctive aspects in youths' lives, including various risk and protective factors in the community, school, family, school, and peer/individual domains. To reduce the burden of the assessed youths and to increase efficiency, the authors would like to create a short adaptive version of the assessment. Due to the multifaceted nature of the original instrument, however, neither unidimensional nor multidimensional IRT-based CAT is a good choice. Unidimensional IRT models do not fit all items. A CAT based on multidimensional IRT models poses prohibitive computational challenges because there are too many dimensions (i.e., 38 scales) and too few items (i.e., 2–10 items) available per dimension.

In this article, a non-IRT-based adaptive test design was presented for the aforementioned risk assessment. The short adaptive test is built based on the *Classification and Regression Trees* (*CART*) *framework* (Breiman et al., 1983), a powerful and increasingly popular machine learning technique (James et al., 2013). This method is nonparametric and does not pose any assumption on dimensionality. It is especially suitable for measurement instruments designed to cover a large number of distinctive areas. In this article, the authors call their design ''tree-based adaptive test.'' Examples of pioneer work that created tree-based adaptive tests are Wainer et al. (1991), Wainer et al. (1992), Yan et al. (2004), and R. D. Gibbons et al. (2013).

In the remainder of the article, the authors review the literature on IRT-based and tree-based adaptive tests. Then, they present the details of their design, which involves advanced techniques to address the imbalanced data problem. With a sample data set of 3,975 subjects, and

using cross-validation, the authors compare the results of eight variations of tree-based adaptive tests and a set of benchmark methods including four variations of random forest and unidimensional IRT scoring. The results suggest that with the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2002), the tree-based adaptive tests were able to achieve the most superior efficiency. The final tree-based adaptive tests reduced the instrument from 173 items to only 3 to 6 items taken by an individual. Following the presentation of these findings, the authors discuss the promise, challenges, and future directions of tree-based adaptive tests.

Literature Review

IRT-Based Adaptive Test Studies

The literature on IRT-based adaptive test (a.k.a., CAT) is abundant and cannot be fully summarized here; hence, the authors only discuss the few studies that explicitly used IRT-based CAT for shortening long measurement instruments. Ware et al. (2003), Hart et al. (2008), and L. E. Gibbons et al. (2011) created CATs based on unidimensional IRT models. In those adaptive tests, computer algorithms select a personalized subset of items for each individual based on IRT. In IRT, a measured trait is conceptualized as a continuous latent interval scale, and the purpose of the measurement is to pinpoint one's location on the scale (a.k.a., the individual's trait level). As the test proceeds, the individual's estimated trait level is constantly updated using statistical estimation procedures based on the item responses provided by the individual to date. Each new item is selected to maximize the IRT model–based statistical information at the individual's provisional trait level. Intuitively, if one answers the first question correctly, the next question should be more challenging to help pinpoint the individual's location on the latent trait scale, while another easy item will not provide much additional information. This technique can result in a significant reduction in total test length by omitting noninformative items, while holding measurement accuracy on par with the full version.

CAT based on multidimensional IRT models has also been proposed to efficiently measure multiple traits at the same time (e.g., Seo & Weiss, 2015; Wang et al., 2019; Yao et al., 2014; Zheng et al., 2013). Such CAT algorithms based on multidimensional IRT models are statistically and computationally intense. The computational intensity further increases dramatically as the number of dimensions increases. Therefore, multidimensional CAT may be feasible for a handful of dimensions (e.g., 2–4) but might not be practical when a large number of traits are being measured.

Tree-Based Adaptive Test Studies

Tree-based adaptive tests have been built for both classification purposes and for predicting a continuous score. Here, the authors begin with explaining the classification type of tree-based adaptive tests. An artificial example is illustrated in Figure 1. Each test-taker starts by responding to the item on the top (root) node, proceeds down node-by-node based on the observed item response data, and is finally classified into one of the several categories at the end (leaf) node. Note that all nodes are spilt two-way, which is an application of the *CART framework* (Breiman et al., 1983) in machine learning. Different test-takers may go down different branches of the tree, which is similar to receiving different items in an IRT-based adaptive test. Note that the items can be dichotomous (e.g., Questions 15 and 3) or polytomous (e.g., Questions 42 and 38).

R. D. Gibbons et al. (2013) developed a tree-based adaptive classification test for the purpose of screening for major depressive disorder. They used the R package **rpart** to implement

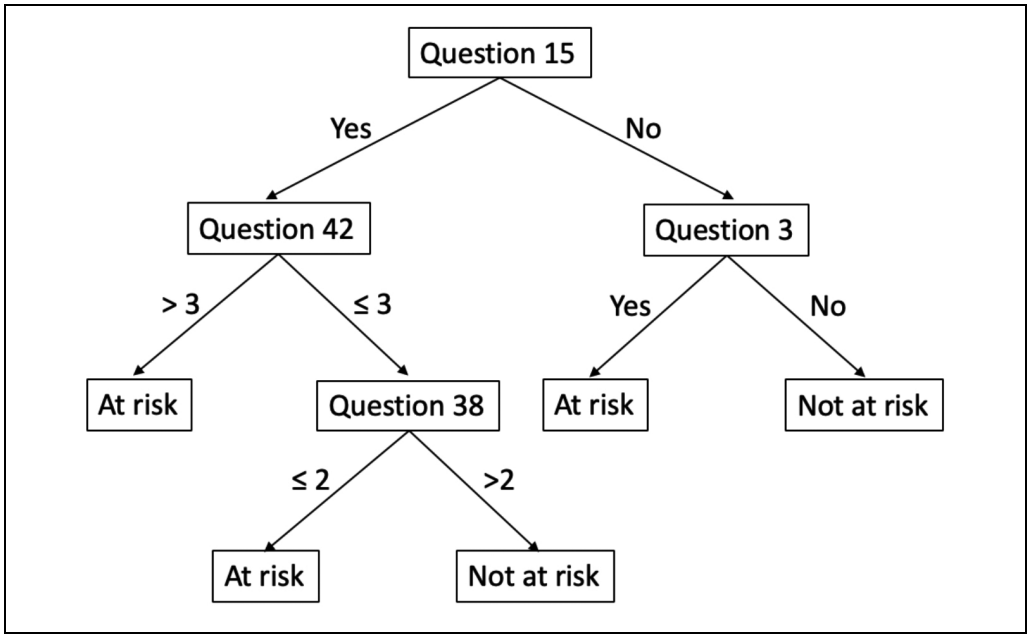

Figure 1. Illustration of a tree-based adaptive screening test.

the CART framework (Breiman et al., 1983) and selected the optimal items from a pool of 88 items to construct the tree. Their final tree-based adaptive test reduced the original instrument of 88 items to an average of 4.2 items per patient, while maintaining a sensitivity/specificity of .95/.87 for predicting the binary outcome.

Tree-based adaptive tests have also been constructed to predict the continuous full-test score. Yan et al. (2004) proposed the use of tree-based adaptive tests to ''predict the observed scores that test-takers would have received if they had taken every item in a reference test or a pool'' (p. 295). Similar to R. D. Gibbons et al. (2013), Yan et al. constructed their tree-based on the CART framework. The outcome of their tree-based adaptive test, instead of classification, is the predicted observed score of the full test. The predicted outcome given by each end node is the average full test observed score of all test-takers in the training sample (i.e., the sample data used to construct the tree) that belong to the node.

Background of the Studied Case: Juvenile Delinquency Risk Assessment

The objective of the assessment is to screen youths for risk of future delinquency, ranging from minor issues such as truancy to serious violence, and then target the identified at risk youth for intervention services so that delinquency is prevented. An instrument was developed that measures relevant risk factors and protective factors surrounding youth in four domains. Table 1 below shows the number of scales, each measuring a risk/protective factor, included in the questionnaire for each domain. Each scale consists of two to 10 questions measured either on Likert-type scales or by yes/no responses, totaling 173 items for the entire questionnaire.

Through a previous research project, complete questionnaire response data from a sample of 3,975 school youth were collected in a Central American country. For every youth in the

Table 1. Numbers of Scales Included in Each Domain.

Scales	Community domain	School domain	Family domain	Peer/individual domain
Risk factor scales	5	4	4	14
Protective factor scales	2	3	2	4

Table 2. Distribution of Outcome Measures in the Sample (N = 3,975).

Engaged in problem behavior in the last six months	Violent Behaviors (%)	Property crime (%)	Gang involvement (%)	Drug sales (%)	Carrying weapons (%)	Truancy (%)
Yes	11.7	20.4	2.3	1.8	2.3	17.0
No	88.3	79.6	97.7	98.2	97.7	83.0

sample, six self-reported delinquency outcomes were collected: violent behaviors, property crime, gang involvement, drug sales, carrying weapons, and truancy. Project participants were asked to report whether they have engaged in particular types of problem behavior in each category. If a youth responds ''yes'' for at least one problem behavior in a category, the respondent was coded as ''yes'' for that outcome category; otherwise ''no.'' For lack of future observed outcome data, these self-reported concurrent delinquency measures serve as the outcome data in the study. The validity of self-reported data has been demonstrated repeatedly in prior criminological studies. Prior research suggests more than 80% convergence between self-reported offending and official reports of arrest (Bersani & Piquero, 2017; Piquero et al., 2014) and has reported a strong and robust relationship between self-reported drug use and urinalyses tests of illicit drugs (Cheon et al., 2018). The distributions of the outcome measures in the sample are given in Table 2 below.

Validation and Scoring of the Scales

Given the structure of the questionnaire (38 scales, 173 items), it is possible to build trees based on either the 173 individual items or the 38 scales. When scales are used as the nodes in the trees, a respondent needs to complete the entire scale (2–10 items) before being routed to the next node. The composite score of the scale is used as the continuous value of the node. To justify the use of scale scores as valid measures, the authors examined the internal consistency reliability of the 38 scales using coefficient ω (McDonald, 1970) and conducted a series of single-factor *confirmatory factor analysis* (CFA), one CFA per scale, to examine factorial validity through model fit statistics. Results show that responses to all the 38 scales were internally consistent with the ω values between .59 and .91 and one scale at .42, and the CFA models retained excellent model fit (root mean square error approximations [RMSEAs] $<$ .083 with one scale at .113, comparative fit indices [CFIs] $>$ .930, Tucker–Lewis indices [TLIs] $>$ .905). Full results are presented in Online Appendix E.

With the support of the above scale validity evidence, the authors created practical scale composite scores based on the average Percent of the Maximum Possible (POMP; P. Cohen et al., 1999) scores. The POMP method was applied because the items include different numbers of response options across the scales. To have all items weight equally in the average score,

all item responses were converted to 0 to 100 using the POMP method. Then, the POMP item scores of each scale were averaged to render the scale composite scores. The protective (the opposite of risk) factor scale scores were reversed to align with the risk factor scales.

Method

Decision Tree and Classification Tree

The application of trees in psychometrics was explained in the ''Literature Review'' section. The authors now discuss the technical aspect of constructing a tree. A critical task when constructing a tree is to identify those attributes that best predict the outcome. In psychometrics, it is natural to use individual item responses as attributes; in the case study, the authors also explored the option of using scale composite scores as attributes. As the first step of building a tree, the attribute (item score or scale composite score) that best predicts the outcome (categorical or continuous) is selected among all available attributes to be the root node. The subsequent nodes are sequentially constructed with other attributes that best enhance the prediction.

A second critical task of tree construction is to determine the branching rules on each node. For a dichotomous item, the two branches are naturally based on the binary response. For a polytomous item or a continuous scale score, however, an optimal cutoff value needs to be determined. Both tasks, the optimal choice of the attribute for a node and the optimal choice of the cutoff value for a given attribute, are typically carried out by optimizing a statistical criterion based on the training sample data.

The most popular criterion for classification trees is the Gini index (James et al., 2013, p. 312). It measures the impurity of a node: A node is purest when the node contains observations from only one class, rendering the minimal values for the criterion; the purity of the node decreases, and the criterion value increases, when the node contains observations from a more diverse group of classes. In this study, the authors used the popular R package **rpart** to build their classification trees, which implements the CART algorithm originally described in the book *Classification and Regression Trees* (Breiman et al., 1983), with the Gini index as the optimized criterion. It is a greedy heuristic algorithm that grows a tree node-by-node by maximizing the purity of the new children nodes at each step (Therneau & Atkinson, 2017).

Random Forest

In practice, when researchers adopt a decision tree approach to solve their problems, they often go one step further to build a so-called *random forest* (Breiman, 2001). Random forest is one of the most popular *ensemble methods* in machine learning. The idea of ensemble methods is to combine the power of a large number of weak learners (e.g., small trees) to enhance the prediction accuracy of the ensemble, especially to overcome the *overfitting* problem typical to single learners (e.g., a single decision tree). The overfitting problem refers to the loss of predicting accuracy for cases outside the existing sample data. In other words, a tree is grown to fit the given data set to the most perfect extent. But the data set is only a sample of the entire population and contains random sampling errors. Thus, the tree that ''overfits'' the given sample usually loses a substantial amount of accuracy when it is used to predict new cases. To overcome this limitation, various ensemble methods have been proposed. One of the most accurate and robust ensemble methods is random forest, which grows a large number of decision trees, each time on a bootstrap sample and restricted to a random small subset of attributes, and uses the majority vote of all trees to predict a new case.

However, random forest, or any other ensemble method, is not appropriate for our problem of building an adaptive test because the goal is to create a single adaptive test that users can administer. That means the authors need a single tree, not a forest. Yet the authors included random forest (implemented by the R package **randomForest**) in their cross-validation study as one of the benchmarks because it has been proven the winner of the competition of popular methods over a large number of problems (e.g., Tan et al., 2006, pp. 293–294).

Imbalanced Data

Imbalanced data are a challenging situation in classification problems. It refers to when some of the classes occur at a very rare rate. The default uses of machine learning classifiers do not do well in predicting rare classes because the majority class data dominate the algorithm, and the accuracy of identifying the rare class cases is not optimized.

The study suffers from the imbalanced data problem. Especially, only 2.3% of the sample reported involved in gangs, 1.8% reported having participated in drug sales, and only 2.3% reported having carried weapons (see Table 2). To enhance the performance of the trees for such imbalanced data, the authors applied weights to the two outcome classes. In the cross-validation study, the authors included three conditions: (a) unweighted; (b) balanced: weighting the two outcome classes in reciprocal proportion to their frequency counts in the sample data, which is equivalent to equalizing the presence of the two classes; and (c) over-weighted: doubling the weights of the minority outcome classes from the second condition.

To enhance the performance of random forest, the authors applied a subsampling method that creates balanced training data. In other words, the data used to train the model are not the entire sample; instead, all minority class cases are included, but only a random subset of the majority class cases are included, where the size of the subset is held equal to the number of minority class cases. The cross-validation study also includes the original random forest method that uses the entire sample data.

Trees With SMOTE Technique

Another method to deal with imbalanced data is the SMOTE (Chawla et al., 2002). As its name suggests, it is an oversampling method designed to combat imbalanced data problem. It creates synthetic samples from the minority class to increase the presence of minority class cases in the data. For each minority class case, the algorithm selects k similar instances based on a distance measure (refer to the k-nearest neighbors algorithm, James et al., 2013, pp. 104–109) and perturbs an instance one attribute at a time by a random amount within the difference to the neighboring instances. The authors used the **SMOTE** function provided in the R package **DMwR** to implement this technique.

In addition to creating synthetic samples for the minority class, the authors also created additional synthetic samples for the majority class using the same method. To build a classification tree for each outcome measure, the authors created a separate SMOTE sample where ~20,000 cases are from the minority class, and another ~20,000 cases are from the majority class. The SMOTE sample size is about 10 times of the original sample size. This approach serves two purposes: (a) balancing the two classes to overcome the imbalanced data problem and (b) using a very large sample size to overcome the overfitting problem of single trees.

Cross-Validation Study

The authors carried out a fivefold cross-validation study to compare the results of different methods. For the tree-based adaptive tests, the authors included (a) unweighted classification tree, (b) balanced-weighted classification tree, (c) overweighted classification tree, and (d) classification tree based on the SMOTE sample. For random forest, the authors included (e) original random forest and (f) balanced random forest. The authors also crossed these methods with two options for the unit for attributes: (a) individual items and (b) scale composite scores.

In a fivefold cross-validation, the authors randomly divide the entire sample into five equal folds (i.e., five equal-sized subsets of the data). Then, the authors take each fold as the testing data, while taking the remaining four folds as the training data. The authors first use the training data to develop the tree or forest. Then, the authors apply the resulting tree or forest to the testing data to assess the classification accuracy, where for each case in the testing data the authors compare the predicted outcome given by the tree or forest against the actual observed outcome. The accuracy measures are averaged across the five iterations. Compared with reporting in-sample accuracy, cross-validation reports out-of-sample accuracy, which is a more truthful estimation of the accuracy of the classifiers when they are used in real practice. For a fair comparison, the same training set and test set were used across the conditions in each iteration. Note that for the condition of classification tree based on the SMOTE sample, the training sets were formed differently by taking a random 80% of the large SMOTE synthetic sample, but the testing set was held the same as that used for all other conditions, which is a 20% subset of the original sample.

IRT Scoring With ROC Optimal Cutoff

In addition, the authors included the results from unidimensional IRT scoring based on the *graded response model* (GRM; Samejima, 1972) and an optimal cutoff value determined by the Youden method (Youden, 1950) based on the *Receiver Operating Characteristic* (ROC) curve. As the authors explained previously, the scales measure distinctive areas of a youth's life and are expected to share too little communality to be adequately modeled by unidimensional IRT models. Yet the authors included this method to provide a comparison between the machine learning methods and IRT methods on the same data.

The unidimensional IRT model extracts a single latent score that reflects the shared, expectedly weak, factor. The authors applied IRT scoring to the same fivefold cross-validation data. Within each iteration, the authors used the R package **mirt** to calibrate GRM item parameters of the 173 items based on the training set. After the initial calibration, the authors excluded the items that failed to render reasonable item parameter estimates (any estimated item parameter with an absolute value greater than 10) and recalibrated the item parameters. Over the five cross-validation iterations, between 12 and 14 items were excluded each time. (Note that this procedure may be regarded as a usual but by no means the best practice of item calibration.) Then, the authors estimated the latent trait score of each respondent in the training set based on the final item parameters using the R package **catR** with the maximum likelihood estimator and infinite scores censored to $[-4, 4]$. These latent trait scores of the training set, together with the corresponding observed binary outcome data, were used to generate an ROC curve and find the optimal cutpoint on the curve.

When a continuous score is used to predict a binary outcome, a cutpoint is needed to generate a classification—a score above the cutpoint leads to the classification into one group, whereas a score below the cutpoint leads to another group. Different choices of the cutpoint results in different levels of sensitivity and specificity (see next section for the definition of the terms). The

		Predicted by the Instrument		
		Positive	Negative	
Observed Outcome	Positive	*A*	*B*	Real Positives = *A*+*B*
	Negative	*C*	*D*	Real Negatives = *C*+*D*
				Grand Total *N* = *A*+*B*+*C*+*D*

Figure 2. A two-by-two contingency table. A, B, C, and D are frequency counts that fall into each cell.

trade-off between sensitivity and specificity as associated with the varying cutpoint can be depicted by an ROC curve. In the study, the authors used the Youden method (Youden, 1950), implemented by the R package **OptimalCutpoints**, to determine the optimal point on an ROC curve where the sum of sensitivity and specificity is maximized. Specifically, the Youden index J, defined in Equation 1 below, is maximized (see next section for the definition of terms).

$$J = \frac{A}{A+B} + \frac{D}{C+D} - 1. \quad (1)$$

This is a commonly used criterion that combines sensitivity and specificity into one index with equal weights.

Then, to evaluate the out-of-sample classification accuracy, latent trait scores of each respondent in the testing set were estimated in the same way as the training set. Those whose scores were greater than the optimal cutpoints were classified as at risk and compared against the observed outcome. The accuracy measures were averaged across the five cross-validation iterations.

Accuracy Criteria

Figure 2 above illustrates a two-by-two contingency table that summarizes the results of a binary classifier.

For regular balanced data, an overall quality index is the *accuracy index*, which is the proportion of correctly classified cases: $(A + D) / N$. In an imbalanced data scenario, however, accuracy is not a useful criterion because it does not reflect the classification accuracy of the minority class cases. As long as the majority class cases are correctly classified, the accuracy index value will be high even if the minority class cases are poorly classified. A useful alternative is Cohen's κ (J. Cohen, 1960), which takes into account the distribution of the observed classes in the data. It is an adequate criterion to reflect classification accuracy in both classes for imbalanced data. Cohen's κ is calculated by the following equation:

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where $p_0 = accuracy = (A + D)/N$, and $p_e = [(A + B)(A + C) + (C + D)(B + D)]/N^2$.

In addition to Cohen's κ, two other useful criteria are class-specific accuracy indices: *sensitivity* and *specificity*. Sensitivity is the proportion of true positives correctly identified as positive by the instrument: $A / (A + B)$. Specificity is the proportion of true negatives correctly identified as negative by the instrument: $D / (C + D)$. In the study, the authors compare Cohen's κ, sensitivity, and specificity for each method.

Results

Cross-Validation Results

Figure 3 presents the classification accuracy results from the cross-validation study. Due to space limits, the table of results in the numeric form is presented in the Online Appendix.

Individual items versus scale composite scores as the attribute unit. Except for IRT scoring, the authors compared the options of using individual items versus scale composite scores as the nodes in the trees. When scales are used as the nodes, each respondent needs to complete the entire scale (2–10 items) at each node. In general, cross-validation results show that the overall classification accuracy (as indicated by Cohen's κ) is similar between the two options, and

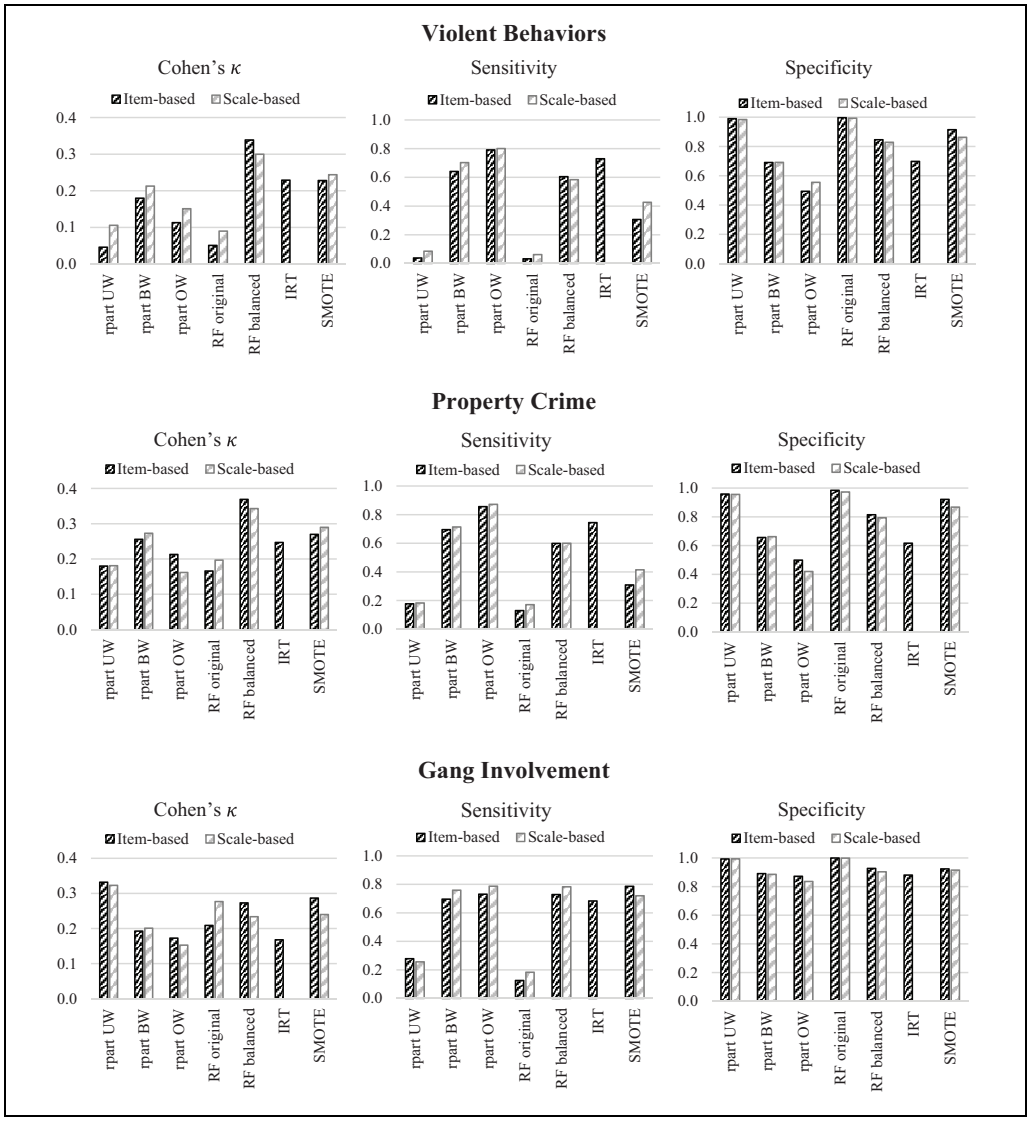

Figure 3. (continued)

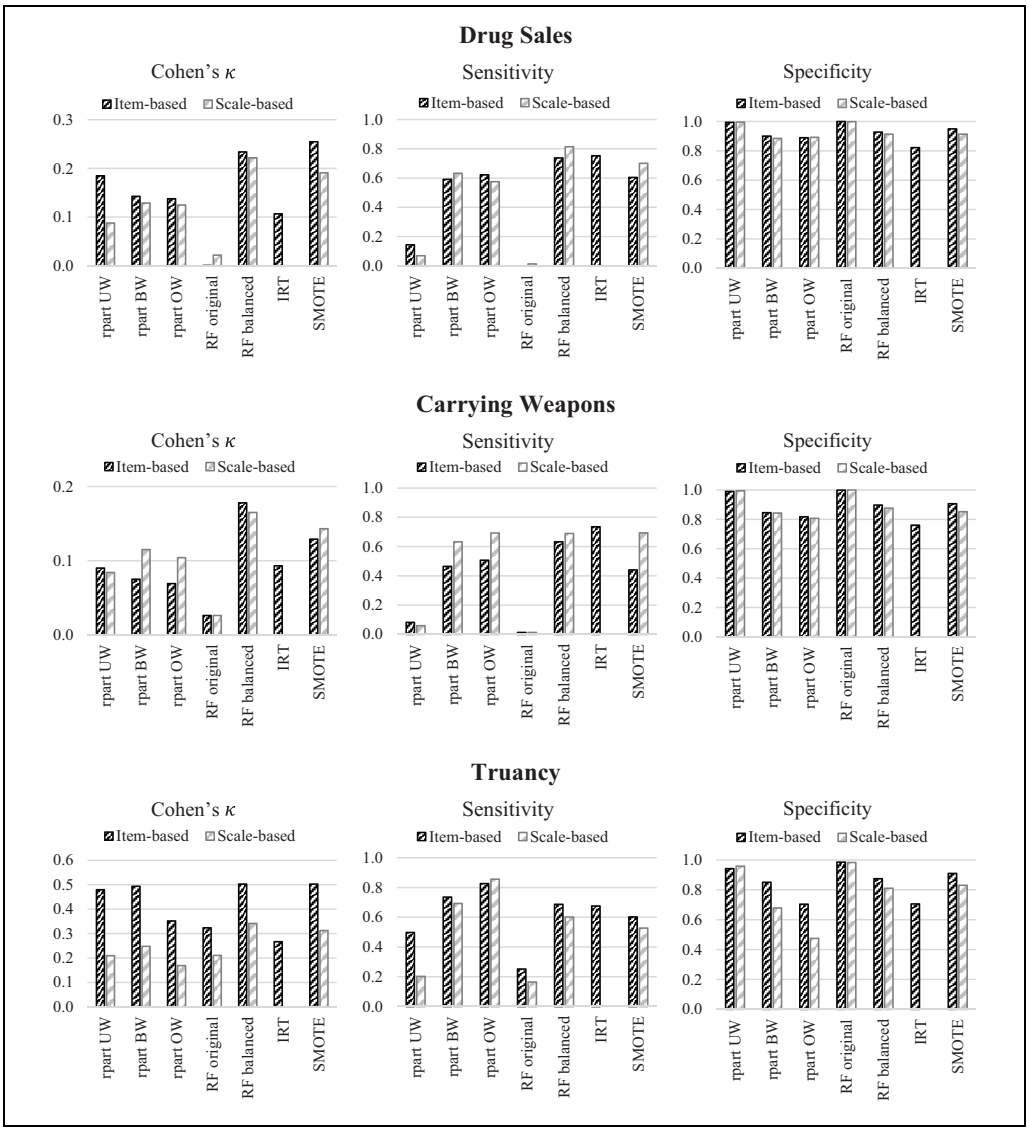

Figure 3. Cohen's κ, sensitivity, and specificity values of compared methods from the cross-validation study.

Note. UW = unweighted; BW = balanced weighted; OW = overweighted; RF = random forest; SMOTE = rpart with Synthetic Minority Oversampling Technique sampling; IRT = item response theory.

there is no clear winner. The only exception is seen for truancy, where item-based results were noticeably better than scale-based results. Given these results and the fact that item-based tree adaptive tests are much shorter than the scale-based tree adaptive tests, item-based methods are more desirable within the context. The authors will provide further discussion in the last section, but *hereafter the authors will focus on examining the results of the item-based conditions.*

Results confirming consequences of imbalanced data. Comparing the seven methods based on individual items, the original random forest yielded the lowest κ values in four of the six outcomes

and very close to the worst condition in another outcome. Given that random forest has been proven the best-performing method for a large number of problems (e.g., Tan et al., 2006, pp. 293–294), this result may seem surprising. However, it confirms the concern regarding imbalanced data. The majority class cases, the negative class in this study, dominated the algorithm and achieved almost perfect specificity (.985~1.000); But the sensitivity values are very low (.000~.252). Similar patterns are seen in the trees with the original rpart algorithm (specificity ranging .942~.994, sensitivity ranging .039~.498), for the same reason. What further confirms the reasoning of the issue are the patterns in the results of the weighted rpart and balanced-sampling random forest algorithms. Figure 3 shows that the weighting and balanced-sampling techniques significantly improved the sensitivity for all six outcomes. Furthermore, for rpart trees, the increase of sensitivity is monotonic from unweighted to balanced-weighted to over-weighting minority cases, which is anticipated based on the reasoning.

Classification tree built with SMOTE sampling technique. The cross-validation study results have demonstrated the effectiveness of the SMOTE technique. With one exception, the SMOTE rpart method yielded the greatest κ value among all compared tree methods. For three of the six outcomes, the SMOTE rpart method yielded a κ value greater than or equal to the benchmark method random forest with balanced sampling. Comparing SMOTE rpart with the other benchmark, IRT scoring with optimal cutoff, except for one outcome where there is a close tie, for all other outcomes, SMOTE rpart yielded noticeably higher κ values. Note that both random forest and IRT scoring rely on all 173 items, while SMOTE rpart, as seen in the next section, only needs to administer between 3 and 6 items per respondent. This result suggests that the tree-based adaptive test with the SMOTE rpart algorithm has remarkable efficiency.

Final Tree-Based Adaptive Tests

After evaluating the out-of-sample classification accuracy with the cross-validation study, the authors then applied the winning tree method, rpart with SMOTE, to the entire sample to create the final tree-based adaptive tests. The final trees are given in Online Appendix B. Each individual starts with the root node on the top and only proceeds following one path based on his or her responses. The depth of the tree is the maximum number of items administered to the subjects. Some people may take the paths that are even shorter. The range of test lengths is only 3~6 items per tree, and the final trees have similar accuracy as the cross-validation results, with κ values ranging .133~.504, sensitivity ranging .309~.791, and specificity ranging .884~.950. Online Appendix C lists the items that are included in each tree. The contents of the items are all highly relevant to the specific outcomes, providing evidence of content validity of these adaptive tests.

Stability of Results Across Varied Samples

One concern about tests built based on data-driven decisions is whether the resulting tree-based adaptive tests will be different when a different sample of data is used to train the trees. In fact, as mentioned in the ‘‘Method’’ section, one of the purposes of applying the SMOTE technique and creating a very large synthetic sample of ~40,000 cases was to overcome overfitting and stabilize the results. To examine the stability of the results, the authors conducted a sensitivity analysis by comparing the final tree created with the entire SMOTE sample for each outcome (as shown in Online Appendix B) with the five trees created by the fivefold cross-validation subsamples. Each cross-validation subsample contained a random subset of 80% of the whole sample. Table 3 summarizes the differences by two metrics: (a) the number of trees (of five)

Table 3. Stability of Results Crossed Varied Samples.

Number of identical trees	Delinquency outcome	% of nodes in final trees altered in subsample trees					
		Subsample 1	Subsample 2	Subsample 3	Subsample 4	Subsample 5	Average (%)
Violent behaviors	2	1/13 (7.7%)	0/13 (0%)	0/13 (0%)	1/13 (7.7%)	3/13 (23.1%)	7.7
Property crime	3	0/11 (0%)	4/11 (36.4%)	4/11 (36.4%)	0/11 (0%)	0/11 (0%)	14.5
Gang involvement	5	0/7 (0%)	0/7 (0%)	0/7 (0%)	0/7 (0%)	0/7 (0%)	0
Drug use	2	1/8 (12.5%)	1/8 (12.5%)	0/8 (0%)	4/8 (50%)	0/8 (0%)	15
Drug sales	5	0/6 (0%)	0/6 (0%)	0/6 (0%)	0/6 (0%)	0/6 (0%)	0
Carrying weapon	4	0/9 (0%)	1/9 (11.1%)	0/9 (0%)	0/9 (0%)	0/9 (0%)	2.2
Truancy	4	0/6 (0%)	2/6 (33.3%)	0/6 (0%)	0/6 (0%)	0/6 (0%)	6.7

identical to the final tree and (b) the percent of nodes in the final tree that are altered in each of the five trees created by random subsamples. As the table shows, the stability of results varied across the seven delinquency outcomes. No difference was observed across samples for two of the seven outcomes, for the rest, the average percent of nodes in the final trees altered by subsamples ranged from 2.2% to 15%. In the authors' view, the results are optimistic and suggest the SMOTE technique was effective in stabilizing the results, although not to a perfect extent. From another perspective, when tests are created based on human expert decisions, many decisions may be different across individual experts as well, resulting in different possible tests. When decisions are made based on data, the key is the quality of the data—a large sample representative of the spectrums of characteristics of the target population will help justify the data-based decisions.

Discussion

In this study, the authors explored applying machine learning methods, specifically decision tree methods, to design short adaptive classification tests. In the past, CATs have been designed based on IRT. Yet IRT may not be appropriate for all testing scenarios. When the measured construct(s) include many distinctive domains, unidimensional IRT models do not fit all items, and a CAT based on multidimensional IRT models poses prohibitive computational challenges. For such a testing scenario, the decision tree framework developed in the machine learning discipline, which is nonparametric and does not require any assumptions about dimensionality, becomes a promising solution for creating short adaptive tests. In their study, the authors used data from a large sample to demonstrate that classification trees, when applied with advanced techniques to address specific characteristics of the problem and data, can offer outstanding classification accuracy and superb efficiency in terms of the reduction in test length. Some limitations of this method include the following.

First, many important design factors of IRT-based CAT have not been discussed in this article. For one, the tree-based adaptive test in this article does not apply to tests where cheating is a concern. The method creates a single tree with a handful of items. If there are incentives for cheating, these items can soon be exposed on the internet. The current design is applicable to settings not under the threat of cheating, such as psychological or medical screening, or low-stakes educational assessment settings. As a future direction, random forest or other ensemble methods, where a large number of trees are created, might be viable options to create tree-based adaptive tests that are protected against cheating. Further studies are encouraged to investigate that possibility. Two other design factors are content balancing and item revision (allowing the test-taker to change answers to items already taken). As a future direction, item-set-based trees, where sets of items are used as the nodes, may be a solution to support content balancing and partial item revision in a similar way to multistage testing (Yan et al., 2014) and on-the-fly multistage testing (Zheng & Chang, 2015).

Second, when the authors apply this method to shorten an existing long measurement instrument, the short adaptive test based on a small set of items loses some of the measurement validity that has been established for the full version, such as content validity and factorial validity (AERA et al., 2014). However, the gain in criterion-related validity and measurement efficiency as demonstrated in this study is substantial. Test developers may be able to reconceptualize the measure and reestablish the content validity for the short adaptive tests. But more importantly, with the advent of the machine learning era, the psychometrics discipline is now at a critical crossroad, demanding the community to initiate more conversation and research on how to best integrate the old and new concepts, frameworks, and techniques to create a new era of psychometrics.

Acknowledgment

The authors thank the editors, the anonymous reviewers, and Dr. Terry Ackerman for their constructive comments and thank Wina Kurniawan for her contribution to data analysis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Yi Zheng https://orcid.org/0000-0003-2671-0820

Supplemental Material

Supplementary material is available for this article online.

Note

1. Examples are the U.S. Armed Service Vocational Aptitude Battery (ASVAB; Sands et al., 1997), the Graduate Management Admission Test (GMAT), and the Graduate Record Examination (GRE).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Bersani, B. E., & Piquero, A. R. (2017). Examining systematic crime reporting bias across three immigrant generations: Prevalence, trends, and divergence in self-reported and official reported arrests. *Journal of Quantitative Criminology*, *33*, 835–857.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1983). *Classification and regression trees*. Wadsworth.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1–20.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cheon, H., Decker, S. H., & Katz, C. M. (2018). Medical marijuana and crime: Substance use and criminal behaviors in a sample of arrestees. *Journal of Drug Issues*, *48*(2), 182–204.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, P., Cohen, J., Aiken, L., & West, S. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346.

Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., & Patrick, D. (2011). Migrating from a legacy fixed-format measure to CAT administration: Calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research*, *20*(9), 1349–1357.

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., . . . Kupfer, D. J. (2013). The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): A screening tool for depression. *Journal of Clinical Psychiatry*, *74*(7), 669–674.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.

Hart, D. L., Wang, Y.-C., Stratford, P. W., & Mioduski, J. E. (2008). Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. *Quality of Life Research*, *17*(8), 1081–1091.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1–21.

Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime & Delinquency*, *51*, 525–555.

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement*, *18*, 1–68.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. American Psychological Association.

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement*, *75*(6), 954–978.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Pearson Education.

Therneau, T. M., & Atkinson, E. J. (2017). *An introduction to recursive partitioning using the RPART routines*. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Wainer, H. (Ed.). (2000). *Computer adaptive testing: A primer*. Lawrence Erlbaum.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*, 243–251.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement*, *28*, 311–324.

Wang, C., Weiss, D. J., & Su, S. (2019). Modeling response time and responses in multidimensional health measurement. *Frontiers in Psychology*, *10*, Article 51.

Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A. S., Dahlöf, C. G. H., Tepper, S., & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, *12*(8), 935–952.

Yan, D., Lewis, C., & Stocking, M. L. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics*, *29*(3), 293–316.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. CRC Press.

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement*, *38*(8), 614–631.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Zheng, Y., Chang, C.-H., & Chang, H.-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, *22*(3), 491–499.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118.