Article

Applied Psychological Measurement
38(1) 3–17
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621613494466
apm.sagepub.com
SAGE

Combining Decision Trees and Stochastic Curtailment for Assessment Length Reduction of Test Batteries Used for Classification

Marjolein Fokkema[1], Niels Smits[1], Henk Kelderman[1], Ingrid V. E. Carlier[2], and Albert M. van Hemert[2]

Abstract

For classification problems in psychology (e.g., clinical diagnosis), batteries of tests are often administered. However, not every test or item may be necessary for accurate classification. In the current article, a combination of classification and regression trees (CART) and stochastic curtailment (SC) is introduced to reduce assessment length of questionnaire batteries. First, the CART algorithm provides relevant subscales and cutoffs needed for accurate classification, in the form of a decision tree. Second, for every subscale and cutoff appearing in the decision tree, SC reduces the number of items needed for accurate classification. This procedure is illustrated by post hoc simulation on a data set of 3,579 patients, to whom the Mood and Anxiety Symptoms Questionnaire (MASQ) was administered. Subscales of the MASQ are used for predicting diagnoses of depression. Results show that CART-SC provided an assessment length reduction of 56%, without loss of accuracy, compared with the more traditional prediction method of performing linear discriminant analysis on subscale scores. CART-SC appears to be an efficient and accurate algorithm for shortening test batteries.

Keywords

classification, test batteries, computerized testing, sequential testing, classification and regression trees, stochastic curtailment, respondent burden, efficiency

In many applied settings in psychology, test batteries are used for classification and selection. For example, in mental health care, a battery of self-report questionnaires may be used for performing clinical diagnosis, or for assigning patients to the right treatment. When a number of questionnaires or tests are administered, assessment length becomes an important consideration.

[1] Vrije Universiteit Amsterdam, Netherlands
[2] Leiden University Medical Center, Leiden, Netherlands

Corresponding Author:
Marjolein Fokkema, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, Amsterdam, 1081BT, Netherlands.
Email: m.fokkema@vu.nl

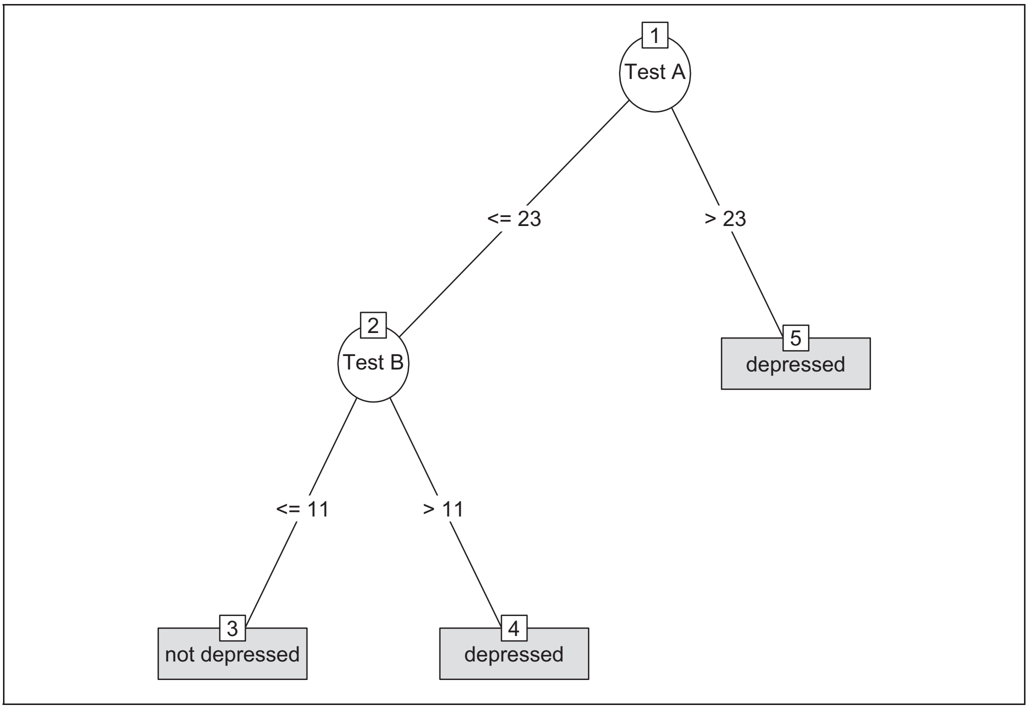

Figure 1. Example of a decision tree for two-stage sequential testing.

Lengthier assessment procedures do not always provide better information, as they may result in adverse side effects as well. For example, lengthy questionnaires are likely to scare off participants, as they have been found to increase attrition (Edwards et al., 2002, 2009; Edwards, Roberts, Sandercock, & Frost, 2004), and have been shown to decrease the quality of responses (Galesic & Bosnjak, 2009; Herzog & Bachman, 1981). Therefore, minimizing the respondent burden while maximizing the efficiency of test batteries is an important prerequisite for accurate classification (Finkelman, Smits, Kim, & Riley, 2012). To improve efficiency, two approaches may be taken: omitting redundant tests from the assessment and omitting redundant items from the tests.

To omit redundant tests from assessment procedures, Cronbach and Gleser (1965) introduced sequential testing. Sequential testing aims to collect new information at every stage of testing, and neglects attributes that are redundant, given previous outcomes. At every stage of sequential testing, a new test is selected, which is most informative for predicting class membership, until a final classification decision can be made. So instead of administering the same sequence of tests to all participants, only those tests that contribute to the classification decision are administered, resulting in a more efficient testing procedure. The stages in a sequential testing procedure may be represented by a decision tree. For example, in Figure 1, the sequential testing procedure for classifying patients as depressed or not is depicted. In this procedure, Test A is administered first: If a patient's score on Test A exceeds a cutoff value of 23, testing can be halted, and the patient can be diagnosed with depression. Only when a patient's score on Test A does not exceed 23, administration of Test B is required to make a final classification decision.

To develop a sequential testing plan, the optimal sequence of tests and their cutoff values have to be derived. As the stages in sequential testing can be represented by a decision tree, classification and regression tree (CART) models provide a natural solution. CART algorithms provide tree-like structures, comparable with Figure 1. In every split, the algorithm creates subgroups for which the distributions of the outcome variable (e.g., classification decision) are most different. Consequently, every split in such a tree provides the current most informative test and cutoff value for making a classification decision.

Reducing the number of tests is not the only way to reduce assessment length. Many methods have been developed to reduce the number of items within a test. Traditionally, test length reduction was aimed at creating fixed-length tests, in which the subset and order of items is determined a priori, before test administration (e.g., Burisch, 1997). Over the last three decades, due to the increasing use of computers in psychological testing, efforts have been aimed at adaptively reducing the number of items, by creating variable length tests. With variable length tests, the subset and order of items are determined online, during test administration. As a result, the number and order of items administered may differ between respondents, as a function of the respondents' relative standing on the attribute being measured. Several authors have shown these variable length tests to outperform fixed-length tests (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Fries, Cella, Rose, Krishnan, & Bruce, 2009; Ware et al., 2003) in terms of accuracy and/or test length reduction. These variable length tests can be aimed at either score estimation or classification.

For score estimation, the most popular method is computerized adaptive testing (CAT; for example, Van der Linden & Glas, 2010; Wainer, 2000). With CAT, after administration of every item, a current estimate of the respondent's ability is made. The most informative item, given the respondent's current ability estimate, is administered next. To allow for estimation of the respondent's ability, CATs are developed using item response theory (IRT) models. Non-IRT adaptive testing methods for score estimation have been developed as well. For example, Yan, Lewis, and Stocking (2004); Ueno and Songmuang (2010); and Riley, Funk, Dennis, Lennox, and Finkelman (2011) have used CART as an item selection algorithm in adaptive testing, and compared its performance with IRT-based item selection. Note that this approach differs from the sequential testing approach as described earlier, as CART was used for item selection in these studies, not for test selection. In all three studies, CART performed equally well or better than IRT, in terms of efficiency and accuracy.

For classification, IRT-based adaptive testing methods have been developed, as well. These methods are referred to as computerized classification testing (CCT; for example, Thompson, 2009, 2011). CCT has been successfully applied in educational settings to test students' mastery in specific domains (e.g., Weiss & Kingsbury, 2005). Several non-IRT based adaptive testing methods for classification have been developed as well. For example, Finkelman, He, Kim, and Lai (2011) and Finkelman et al. (2012) recently introduced stochastic curtailment (SC) as an algorithm for improving test efficiency in classification. With SC, items are presented in the same order as in the full-length test, but testing is halted if the remaining items are unlikely to change the final classification decision. Finkelman et al. (2012) found SC to compare quite favorably with assessment by CAT, in terms of efficiency and accuracy of a mental health self-report questionnaire.

In what follows, the authors introduce a procedure for efficient administration of multiple tests or questionnaires for classification and selection, by combining CART and SC. CART will be used to derive the optimal order and cutoff values for test administration and SC to allow for early stopping of item administration within a test. This will result in a reduction of assessment length on two levels: the number of tests in the battery and the number of items in the tests. In the remainder of the introduction, CART, SC, and the combined method (CART-SC) are described in more detail. In the method and results section, an illustration of CART-SC is presented by performing a post hoc simulation study on a real data set. The data set consists of item scores on subscales of a self-report questionnaire covering symptoms of depression, anxiety, and psychological distress. These self-report data are used to predict clinical diagnoses of depression. In the discussion, the findings are summarized, comments on the method are made, and directions for future research are described.

CART

The first tree building algorithm as a tool for data analysis was proposed by Morgan and Sonquist (1963). However, the most popular tree building algorithm was introduced by Breiman, Friedman, Olshen, and Stone (1984), who referred to it as CART. Many authors have suggested refinements or adaptations since, but these can all be seen as special cases of the same algorithm (see also Hothorn, Hornik, & Zeileis, 2006). CART is a nonparametric algorithm, as it makes no assumption about a data-generating function. Using predictor variables, the CART algorithm recursively partitions observations into increasingly smaller subgroups, whose members are increasingly similar with respect to the outcome variable. For qualitative outcome variables, the resulting tree is a classification tree, and for quantitative outcomes, the resulting tree is a regression tree. In the current article, CART will only be used for classification, so the discussion will focus on classification trees. Partitions, or splits, are made using one predictor variable at a time: In every node, the algorithm selects the variable and splitting point that separate the observations into subsets for which the distributions of the outcome variable are most different. The CART algorithm produces binary splits only, that is, the observations are partitioned into two subgroups at every split. The result is a decision tree consisting of branches and nodes. This tree can be used for prediction, by ‘‘dropping’’ new observations down the tree. The distribution of the outcome variable of the training data in the final node determines the prediction for the new observation. For example, the largest class among the training observations in the final node may be used for predicting the class of a new observation.

SC

Finkelman et al. (2011, 2012) introduced SC for shortening questionnaires used for classification, but SC has been applied before in, for example, clinical trial monitoring (Davis & Hardy, 1994). Traditionally, a trial is terminated only when the predetermined number of participants has been included. With SC, the probability of rejecting the null hypothesis at the end of a trial, given the current observations, is calculated. If the probability of rejecting the null hypothesis at the end of a trial is sufficiently high or low, the trial can be stopped early. This can be readily translated into classification using questionnaires: Traditionally, all items of a scale are administered first, and then a test score is calculated and compared with a predetermined cutoff value. With SC, after administration of an item, a cumulative score is calculated, and the probability of obtaining a test score exceeding the cutoff at the end of the questionnaire, given the cumulative score, is calculated. This probability can be calculated by empirical proportions, obtained from a training data set. In the current study, this empirical approach is applied, but Finkelman et al. (2012) have proposed a model-based variation of SC as well: logistic SC, in which the probability of exceeding the cutoff value is calculated by means of logistic regression (LR). As long as the probability of exceeding the cutoff value is below some predetermined threshold, testing continues; if not, testing is halted and a classification decision is made. SC (logistic and empirical) requires a training data set to determine the probability of obtaining a final positive (‘‘at risk’’) or negative (‘‘not at risk’’) classification decision. Applying empirical SC for classification using a single scale consists of the following steps (Finkelman et al., 2012):

1. The cutoff value for classifying observations as ‘‘at risk’’ (X^*) is determined.
2. The threshold for the probability of making an incorrect risk classification (γ) is chosen by the user.
3. The learning data set is split into two parts: T^+, containing item scores of all participants ‘‘at risk’’ (i.e., with a test score equal to or exceeding X^*), and T^-, containing item scores of all participants ‘‘not at risk’’ (i.e., with a test score less than X^*).
4. For every new (i.e., not in the learning data set) respondent, a cumulative score is calculated after administration of every item k. All answers to items $k + 1$ through N are taken from the T^+ data set and appended to the cumulative score at item k. The proportion of the resulting test scores exceeding X^* is denoted by $\hat{P}_k^+$.
5. Similarly, all answers to items $k + 1$ through N are taken from the T^- data set and appended to the cumulative score at item k. The proportion of resulting test scores exceeding X^* is denoted by $\hat{P}_k^-$.
6. When $\hat{P}_k^+$ and $\hat{P}_k^-$ are both $\geq \gamma$, testing is halted, and an ‘‘at risk’’ classification is made for the respondent. When $\hat{P}_k^+$ and $\hat{P}_k^-$ are both $\leq (1 - \gamma)$, testing is halted, and a ‘‘not at risk’’ classification is made for the respondent. Otherwise, the next item is administered, and Steps 4 through 6 are repeated.

Note that in Steps 2 and 6 of the algorithm, Finkelman et al. (2012) used two γ values instead of one. This allows for specification of different thresholds for the probability of making an incorrect decision for ‘‘at risk’’ and ‘‘not at risk’’ classifications. In the application of SC in the current article, the same threshold is used for both probabilities, so one γ value suffices.

Applying the SC algorithm may provide substantial reductions in assessment length: Finkelman et al. (2011) showed that SC could reduce the average test length of a health questionnaire by 42%, while the curtailed and full-length instruments showed identical diagnostic accuracy. Likewise, Finkelman et al. (2012) applied curtailment and SC to shorten the Center for Epidemiological Studies–Depression Scale (CES-D; Radloff, 1977) and showed that test length could be reduced by up to 23% on average, with identical diagnostic accuracy, compared with administration of the full-length CES-D.

CART-SC

As noted earlier, sequential testing using CART may provide a powerful tool for reducing the number of tests to be administered for classification. A classification tree can be built using test scores on several scales as inputs and classifications as outputs. By using this tree to guide test administration, only those scales that are necessary for classification are administered, thereby reducing the average total test length. In addition, the classification tree would provide optimal cutoff values for every scale. In turn, these cutoff values can be used to apply SC in every node of the tree, resulting in an additional reduction of items administered within each scale.

Like CART and SC, the CART-SC algorithm consists of two parts: calibration and application. For calibration, a classification tree is grown using all observations in the training data set. Subsequently, for every node (cutoff value) in the tree, a T^+ and T^- dataset is created using all training observations. For application of CART-SC, the first node of the classification tree is used to determine the first scale to be administered to new participants and the cutoff value to be used. Items of this scale are administered until a classification decision can be made according to the last step of the SC algorithm. Finkelman et al. (2012) referred to the decision in this step as an ‘‘at risk’’ or ‘‘not at risk’’ decision, because in their algorithm, only one scale is administered for making the final classification decision. In the CART-SC algorithm, several scales are administered before a final classification decision can be made, so these decisions will be referred to as ''above cutoff'' and ''below cutoff'' decisions. In case of an ''above cutoff'' decision, the next node on the right determines the next scale and next cutoff value to be used for SC. In case of a ''below cutoff'' decision, the next node on the left determines the next scale and the next cutoff value to be used for SC. This process continues until the respondent reaches a final node and the final classification decision can be made.

In summary, CART-SC may provide a substantial reduction in assessment length at two levels: a reduction in the number of tests and in the number of items. In what follows, the potential accuracy and reduction in assessment length of CART-SC will be evaluated by means of post hoc simulation on a real data set.

Method

Data set

Participants. The data set consisted of data points of 3,597 participants, of which 36.8% was male. Mean age was 38.8 years (*SD* = 13.22, range = 17-91). Participants were outpatients at one of three outpatient centers of the Psychiatric Regional Mental Health Care Centers Rivierduinen in Leiden, the Netherlands. They were referred by their general practitioner for a potential mood, anxiety, or somatoform disorder. Of all participants, 46.4% was diagnosed with a current depressive or dysthymic disorder, 43.2% with a current anxiety disorder, and 16.6% with a current somatoform disorder. About one in four participants (26.8%) had comorbid disorders, and about one in four participants (22.9%) did not have any depressive, anxiety, or somatoform disorder. Further details of the sample have been described in Smits, Zitman, Cuijpers, Den Hollander-Gijsman, and Carlier (2012).

Mini-International Neuropsychiatric Interview (MINI). Depressive disorder diagnoses according to the Dutch translation of MINI (Sheehan et al., 1998; Van Vliet & De Beurs, 2007) were used as the gold standard or criterion classification. The MINI is a semistructured psychiatric interview for clinical diagnosis of mental disorders, according to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) and *International Classification of Diseases* (ICD-10; World Health Organization, 1993) criteria. It has been found to have good interrater and retest reliability (Sheehan et al., 1997). All interviews were carried out by a research assistant (a psychiatric nurse or a psychologist). In the current study, participants with a major depressive or dysthymic disorder were labeled as suffering from depressive disorder.

Mood and Anxiety Symptoms Questionnaire (MASQ). Clark and Watson (1991) proposed a tripartite model of anxiety and depression, grouping symptoms into three categories: nonspecific symptoms of general distress, symptoms of anxiety, and symptoms of depression. Based on this model, the MASQ was developed (Watson & Clark, 1991). The MASQ consists of 5 subscales, covering symptoms of anhedonic depression (AD; 22 items), anxious arousal (AA; 17 items), general distress: depression (GDD; 12 items), general distress: anxiety (GDA; 11 items), and general distress: mixed (GDM; 15 items). The AD and AA subscales are indicators of symptoms specific to depression and anxiety, respectively. Internal consistency estimates for the MASQ subscales indicate good reliability: Watson et al. (1995) reported $\alpha \geq .78$ for all subscales, and Wardenaar et al. (2010) reported $\alpha \geq .93$ for the GDM, AD, and AA subscales.

Several authors have studied the predictive accuracy of the MASQ. Bredemeier et al. (2010) reported sensitivity and specificity of .80 and higher for using the AD subscale with a single cutoff for diagnosing depressive disorders. Geisser, Cano, and Foran (2006) performed linear discriminant analysis (LDA) on the MASQ subscale scores and reported sensitivity of .69 and specificity of .80 for depressive disorders. More modest results were reported by Boschen and Oei (2007), who applied LR to predict depressive and anxiety disorders using the MASQ subscales, and reported sensitivities of about .40 and specificities of about .80. In addition, De Beurs, Den Hollander-Gijsman, Helmich, and Zitman (2007) reported good validity for the Dutch translation of the MASQ.

Simulation Design

Cross-validation. As described by, for example, Hastie, Tibshirani, and Friedman (2009), using the same data for calibration and evaluation of a model results in overly optimistic estimates of performance. Therefore, the data set was randomly split in two parts: a training set (n = 1,799) for calibration of the models and a test set (n = 1,798) for application and evaluation of the models.

Classification tree. First, a classification tree was built using all training observations to determine the sequential testing plan. This classification tree was built using the recursive binary partitioning algorithm of Hothorn et al. (2006) consisting of three steps: In the first step, in a given node, a global hypothesis of independence between any of the predictor variables and the response variable is tested. If this hypothesis is rejected, the predictor variable with the strongest association to the response variable is selected. For evaluation of the global hypothesis of independence, parameter α (the probability of falsely rejecting the independence hypothesis in each node) has to be specified. In the second step, for the predictor variable selected in Step 1, a split is made on the value that separates the observations into two subgroups that are most different with respect to the outcome variable. In the third step, Steps 1 and 2 are repeated in each of the subgroups. The algorithm stops when the global hypothesis of independence can no longer be rejected, in Step 1.

In the current study, the predictor variables for building the classification tree were subscale scores on the MASQ. The response variable consisted of diagnoses on depressive disorders according to the MINI. The misclassification costs used for building the tree were equal for false positives and false negatives, and α was set to .05.

To evaluate the accuracy and efficiency of sequential testing by means of CART, the observations in the test set were dropped down the classification tree, and the number of items administered and classification decisions were collected.

Empirical SC. For calibration and application of CART-SC, T^+ and T^- data sets were created using the complete training set for every node in the classification tree. Subsequently, the observations in the test set were dropped down the tree, and for every observation, SC was simulated in every node the observation passed through. The threshold for making an incorrect decision, γ, was set to .05. As the same subscale may appear in multiple nodes of the same tree (with different cutoff points), all item scores are collected during application of the CART-SC algorithm. When the same scale appears in a node further down the tree, the curtailment procedure is first applied to the items already administered, and further items of the scale are administered only if necessary. For every respondent, the number of items administered and the classification decisions were collected to evaluate the accuracy and efficiency of CART-SC.

For the CART and CART-SC simulations, the original item ordering within every scale was preserved. As suggested by Finkelman et al. (2011, 2012), changing the item order for SC, by starting with the most informative items within a scale, may result in a further reduction of assessment length. To test whether changing the item order may further improve performance of CART-SC, CART-SC was applied with two additional item orders. First, items were ordered by item-total correlations within every subscale, starting with the items showing the highest correlations to the subscale score. This approach assumes the most informative items to provide most information about the subscale scores. Second, items were ordered by entry order in a forward stepwise LR model, fit within every subscale, with depression diagnosis as outcome variable. This approach assumes the most informative items to provide most information about the final classification.

Assessment of Performance

Performance of CART and CART-SC in terms of accuracy and efficiency was compared with the performance of two standard classification methods: LDA and LR, using all five subscale scores of the MASQ. LDA and LR have been shown to perform well on diverse sets of classification tasks (e.g., Hastie et al., 2009; Michie & Taylor, 1994; Press & Wilson, 1978). As such, they provide a benchmark for evaluating the accuracy and efficiency of CART-SC. As LDA and LR differ in their assumptions, one method may be preferable over the other, depending on the distribution of the predictor variables (Press & Wilson, 1978). Therefore, both methods were applied to the data set to minimize the effects of distributional assumptions on the performance of the reference classifier.

In addition, the performance of CART and CART-SC was compared with that of a fixed-length assessment length reduction method. By means of receiver operating characteristic (ROC) analysis, the best subscale and cutoff value for predicting depression diagnoses were selected.

Accuracy was evaluated by calculating correct classification rates ([true positives + true negatives] / total), sensitivity (true positives / [true positives + false negatives]), and specificity (true negatives / [true negatives + false positives]) for every method. Efficiency was evaluated by calculating the mean, the standard deviation, and the median of the total number of items administered. Accuracy and efficiency were calculated separately for every node of the classification tree, as well.

Software

R (R Development Core Team, 2010) was used for all analyses. For building the classification tree, the party package (Hothorn, Hornik, Strobl, & Zeileis, 2012) was used, which provides an implementation of the CART algorithm (Breiman et al., 1984) described earlier. The default settings of the package were used. For LDA, the MASS package (Venables & Ripley, 2002) was used. For ROC analysis, the ROCS (Sing, Sander, Beerenwinkel, & Lengauer, 2005) was used. A custom function for SC by empirical proportions was written in R, following the procedure of Finkelman et al. (2012), which was also described previously.

Results

A classification tree for depression classification was built using the training observations; the resulting tree is presented in Figure 2. As can be seen in Figure 2, the first split was made on the AD subscale. In every branch, an additional split was made on the AD subscale farther down the tree. For the most severely depressed subgroup (branch most to the right in Figure 2, containing observations with AD score >87), administration of the AD subscale was sufficient for classification. For all other subgroups, administration of an additional subscale (GDD, GDM, or GDA) was necessary to make the final classification decision. The AA subscale did not appear in the tree, indicating that this subscale was redundant for depression classification.

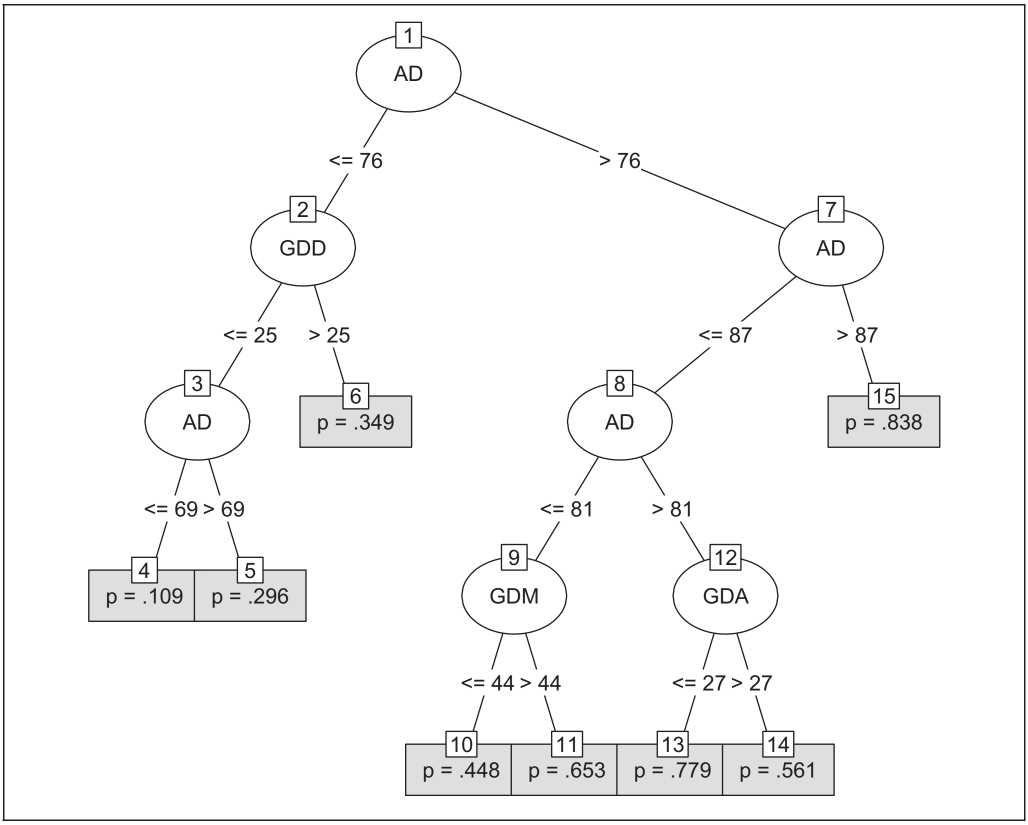

Figure 2. Tree for classification of depressive disorders, with training set prevalences in final nodes.
Note: AD = anhedonic depression; GDD = general distress: depression; GDM = general distress: mixed; GDA = general distress: anxiety.

To provide a benchmark for evaluating the performance of CART and CART-SC, ROC, LDA, and LR analyses were performed on the training data set. ROC analysis of the training data showed the AD subscale to discriminate best between depressed and nondepressed participants, with an area under the curve of .83. Maximum specificity and sensitivity for the AD scale was observed for a cutoff value of 77, in the training data set. Using the AD subscale with a cutoff value of 77 in the test data set resulted in a fixed assessment length of 22 items and a correct classification rate of .73.

The accuracy of LDA and LR were very similar, therefore, only the results of LDA will be discussed. Absolute values of standardized coefficients for the linear discriminant function ranged from .31 to .75, with exception of the standardized coefficient for the AA scale, with an absolute value of .06. This indicates that the AA subscale did not contribute to the prediction of LDA. Therefore, the LDA model was rebuilt, using only the AD, GDD, GDA, and GDM subscales. Using this model for prediction of test observations resulted in a fixed assessment length of 60 items and a correct classification rate of .74 (Table 1).

To evaluate the performance of CART, observations in the test data set were dropped down the classification tree, which resulted in an average assessment length of 31.06, a reduction of 48% compared with prediction by LDA using four subscales (Table 1). The distribution of assessment length for CART showed negative skewness, with higher median than mean

Table 1. Assessment Length and Classification Accuracy in Test Data set (*N* = 1,798).

	LDA	CART	CART-SC
Average assessment length	60.00	31.06	26.50
SD of assessment length	—	1.22	5.46
Median assessment length	60	34.00	27.00
Proportion curtailed[a]	—	—	0.92
Correct classification rate[b]	0.74	0.76	0.76
Specificity[b]	0.75	0.71	0.71
Sensitivity[b]	0.72	0.81	0.81

Note: LDA = linear discriminant analysis; CART = classification and regression tree; CART-SC = classification and regression tree with stochastic curtailment.

[a]Denotes the proportion of observations in the test set, for which at least 1 of the subscales was curtailed.

[b]Accuracy figures are based on test data set, and were calculated with respect to the relevant MINI diagnoses.

(Table 1). The prediction accuracy of the tree was similar to that of LDA. The correct classification rate for CART on the test data was .76, indicating that it performed slightly better than did LDA (Table 1). Specificity of CART was slightly lower than that of LDA (.71 and .75, respectively), whereas sensitivity of CART was notably higher than that of LDA (.81 and .72, respectively).

Applying curtailment in every node of the CART tree resulted in an average assessment length reduction of 15% compared with sequential testing with CART only, and 56% compared with LDA using four subscale scores (Table 1). The distribution of assessment length for CART-SC was symmetrically distributed (Table 1). Predictions for CART-SC were identical to those of CART: Subjects ended up in the same nodes of the tree, with and without curtailment. There was, however, one exception to this rule: One participant ended up in Node 5 in the curtailed tree, but would have ended up in Node 4 in the original tree, due to an inconsistent response pattern. However, this did not change the final classification (no depressive disorder) for this participant.

Changing the item order within every subscale did not result in further reductions, but in minimal increases in average assessment length. Ordering items within every subscale by item-total correlations, and starting every subscale with the items showing the highest correlation to the subscale score, yielded an increase in average assessment length of .13 items. Ordering items within every subscale by entry order in a forward stepwise logistic regression yielded an increase in average assessment length of .08 items. As these changes in assessment length are minimal, they are not presented in Table 1.

For every final node in the classification tree, predictive accuracy and summary statistics for the test length distributions of CART and CART-SC are presented in Table 2. Prevalence (the proportion of participants diagnosed with depressive disorder) in every node was quite similar in training and test data sets, indicating that generalization error is rather small (Table 2). Largest assessment length reductions due to the application of SC were found in Nodes 4 and 6. These were the first and third largest nodes for the test data set and showed assessment length reductions of about 22%.

Predictive accuracy showed some variation across the final nodes of the classification tree (Table 2). Highest predictive accuracy was observed in the nodes with the largest number of observations, Nodes 4 and 15. These two nodes comprised about 25% of all observations each and showed predictive accuracy in the test data of .87 and .79, respectively. Lowest predictive accuracy was observed in Node 10, one of the smaller final nodes, where the proportion of

Table 2. Summary Statistics for Distributions in Final Nodes of the Classification Tree.

						Assessment length		
	Number of observations		Prevalence			CART	CART-SC	
Node	Training set	Test set	Training set	Test set	Accuracy[a]	Average	Average	*SD*
4	488	496	.11	.13	.87	34.00	26.47	3.98
5	108	125	.30	.29	.71	34.00	32.44	1.18
6	278	285	.35	.34	.66	34.00	26.84	3.75
10	105	127	.45	.43	.57	37.00	33.59	1.84
11	72	58	.65	.62	.62	37.00	35.28	1.69
13	131	137	.78	.63	.63	33.00	31.08	1.24
14	123	104	.56	.65	.65	33.00	30.56	1.59
15	494	466	.84	.79	.79	22.00	19.46	1.56

Note: CART = classification and regression tree; CART-SC = classification and regression tree with stochastic curtailment.

[a] Accuracy equals prevalence for cases not classified as depressed and (1 − prevalence) for cases classified as depressed.

correctly classified observations in the test data set was .57 (Table 2). The characteristics of the observations in this node were inspected, as this may provide an explanation for its relatively low classification accuracy. The observations in this node showed above-average AD subscale scores: AD scores in Node 10 ranged from 77 to 81, whereas the overall training set mean was 75.16. At the same time, the observations in this node showed relatively low GDM subscale scores: GDM scores in this node were ≤ 44, whereas the overall training set mean was 40.76. In other words, this group reported relatively high levels of depressive symptomatology but moderate levels of general distress: a somewhat ambiguous profile, which may have resulted in low predictive accuracy. Overall, the variations in predictive accuracy across the final nodes of the classification tree were not specific to CART. The accuracies for LDA and LR in the subgroups defined by the final nodes of the classification tree were equal to, or slightly lower than, that of CART.

Discussion

The results of the post hoc simulation show that the CART-SC algorithm provided a substantial reduction in assessment length. CART-SC provided an average reduction in assessment length of 61% compared with administration of the complete scale, a reduction of 48% compared with LDA on four subscale scores, and a reduction of 15% compared with CART only. At the same time, the classification accuracy of the curtailed tree proved to be identical to that of the uncurtailed tree and even slightly better than that of competitive classification methods (LDA and LR). These findings indicate that CART-SC is an efficient and accurate sequential testing algorithm that may prove useful for shortening test batteries for classification problems in psychology.

The current study provides realistic estimates of accuracy and reductions in respondent burden of CART-SC, as cross-validation was used to estimate classification accuracy of the method. At the same time, the simulation was based on a real data set, which increases confidence in the accuracy and reductions in respondent burden that can be obtained by using CART-SC in applied settings.

As the simulation shows, CART-SC may substantially reduce assessment length in mental health care settings. In many other areas of applied psychology, CART-SC may prove useful in reducing assessment length, as well. The efficiency of selection procedures in, for example, school psychology or personnel selection may be reduced substantially by application of CART-SC. In addition to shortening self-report questionnaires, CART-SC may also be used for shortening other types of tests used for classification and selection, such as mastery tests or neuropsychological test batteries. CART-SC may prove especially useful in settings where tests or questionnaires are already administered by computer in, for example, Internet therapies (computerized self-help interventions delivered by the Internet) or routine outcome monitoring (repeated assessments for monitoring therapy effects in clinical practice; for example, De Beurs et al., 2011). Moreover, it is expected that, in practice, a shorter assessment length will yield better quality data (Galesic & Bosnjak, 2009; Herzog & Bachman, 1981).

As a sequential strategy for assessment length reduction, CART-SC provides some advantages over adaptive testing strategies. First, CART-SC offers test results in the original metric of the scale, which may ease interpretation for psychologists working in applied settings. Second, with CART-SC, the item order of the original instrument can be retained, as the results indicate that item order makes little difference for the performance of CART-SC, in terms of test length and accuracy. In post hoc simulation studies of dynamic adaptive tests, item order is altered after administration of the tests. In practice, changing the item order may result in changes in response patterns, due to item order effects (e.g., McFarland, 1981). Therefore, the test length reduction and accuracy estimates obtained in post hoc simulations may differ from the test length and accuracy that would have been obtained by actual application of the adaptive testing algorithm. For CART-SC, the original item ordering can be preserved in simulation and application of the algorithm, and estimates of performance obtained in post hoc simulation may be better generalizable to future applications of the algorithm.

It should be noted that although CART trees provide intuitively appealing structures, they should not be used as a basis for substantial interpretation. Small changes in the data can result in quite different tree structures, and two rather different tree structures may show equal prediction accuracy on the same data set (Hastie et al., 2009; Hothorn et al., 2006). Nevertheless, inspection of a tree structure may provide valuable information about the relative certainty of the classification decision for a given participant. For participants in nodes with relatively low predictive accuracy, the classification decision may be relatively uncertain and additional testing or interviewing may be necessary. Or the tree structure may be used to indicate potential further reductions in assessment length, by collapsing branches ending in final nodes with the same classification decision. However, these final nodes may be retained to provide information on the certainty of the classification decision made.

Several authors have noted that a potential pitfall of the CART methodology is overfitting: A decision tree may adapt to the idiosyncrasies of the data set too much (e.g., Hand, 2006; Hastie et al., 2009). In the current study, the statistical hypothesis testing approach of Hothorn et al. (2006) was used to counter overfitting, and the data were divided into training and test sets to obtain a realistic estimate of the generalization error. The prevalences in the final nodes of the classification tree in training and test data sets indicated some generalization error, which may have been the result of overfitting. However, classification accuracy was not influenced much, as the classifications were correct for the majority of test observations in every final node.

A potential downside of SC may be that its predictive accuracy is determined by the predictive accuracy of the test score of the full-length test. Relatedly, all item scores are weighed equally and assumed to be equally important for determining the final classification. Items contributing little to the predictive accuracy that may be filtered out by adaptive testing algorithms, may still be administered to many respondents when SC is applied, depending on the position of the item within the test.

In the current study, only empirical SC has been used for shortening assessment length. As suggested by Finkelman et al. (2012), logistic SC is a more rigorous curtailment algorithm, which may result in shorter average test lengths and comparable accuracy, if γ values are set high enough. Therefore, in further studies, the performance of logistic SC within a classification tree may be explored. Another direction for future research is the application of a unidimensional classification CAT in every node of a classification tree: a CART-CAT, instead of a CART-SC procedure.

In short, the current study has shown CART-SC to be an efficient and accurate method for reducing assessment length in classification problems. It may prove useful in many applications in psychology in which several subscales or tests are administered for selection and classification.

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.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Boschen, M., & Oei, T. (2007). Discriminant validity of the MASQ in a clinical sample. *Psychiatry Research*, *150*, 163-171.

Bredemeier, K., Spielberg, J., Silton, R., Berenbaum, H., Heller, W., & Miller, G. (2010). Screening for depressive disorders using the MASQ Anhedonic Depression Scale: A receiver-operator characteristic analysis. *Psychological Assessment*, *22*, 702-710.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. New York, NY: Wadsworth.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, *11*, 303-315.

Choi, S., Reise, S., Pilkonis, P., Hays, R., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*, 125-136.

Clark, L., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, *100*, 316-336.

Cronbach, L., & Gleser, G. (1965). *Psychological tests and personnel decisions*. Champaign: University of Illinois Press.

Davis, B., & Hardy, R. (1994). Data monitoring in clinical trials: The case for stochastic curtailment. *Journal of Clinical Epidemiology*, *47*, 1033-1042.

De Beurs, E., Den Hollander-Gijsman, M., Helmich, S., & Zitman, F. (2007). The tripartite model for assessing symptoms of anxiety and depression: Psychometrics of the Dutch version of the Mood and Anxiety Symptoms Questionnaire. *Behaviour Research and Therapy*, *45*, 1609-1617.

De Beurs, E., Den Hollander-Gijsman, M., Van Rood, Y., Van der Wee, N., Giltay, E., Van Noorden, M., . . . Zitman, F. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*, 1-12.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing response rates to postal questionnaires: Systematic review. *British Medical Journal*, *324*, 1183-1185.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Wentz, R., Kwan, I., . . . Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires ([Review]). *Cochrane Collaboration. Database of Systematic Reviews, 2010*(3), MR000008. doi:10.1002/14651858.MR000008.pub4

Edwards, P., Roberts, I., Sandercock, P., & Frost, C. (2004). Follow-up by mail in clinical trials: Does questionnaire length matter?*Controlled Clinical Trials*, *25*, 31-52.

Finkelman, M., He, Y., Kim, W., & Lai, A. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine*, *30*, 1989-2004.

Finkelman, M., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement*, *36*, 632-658.

Fries, J., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology*, *36*, 2061-2066.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349-360.

Geisser, M., Cano, A., & Foran, H. (2006). Psychometric properties of the Mood and Anxiety Symptom Questionnaire in patients with chronic pain. *Clinical Journal of Pain*, *22*, 1-9.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*, 1-14.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer-Verlag.

Herzog, A., & Bachman, J. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*, 549-559.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2012). party: A laboratory for recursive partytioning (Version 1.0-2.) [Computer software manual]. Available from http://cran.r-project.org/

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651-674.

McFarland, S. (1981). Effects of question order on survey responses. *Public Opinion Quarterly*, *45*, 208-215.

Michie, S. D., & Taylor, C. D. (1994). *Machine learning, neural and statistical classification*. London, England: Ellis Horwood.

Morgan, J., & Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*, 415-434.

Press, S., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*, 699-705.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

R Development Core Team. (2010). *R: A language and environment for statistical computing* [Computer software manual]. Available from http://www.R-project.org

Riley, B., Funk, R., Dennis, M., Lennox, R., & Finkelman, M. (2011, October). *The use of decision trees for adaptive item selection and score estimation*. Paper presented at the Annual Conference of the International Association for Computerized Adaptive Testing, Pacific Groove, CA.

Sheehan, D., Lecrubier, Y., Harnett Sheehan, K., Janavs, J., Weiller, E., Keskiner, A., . . . Dunbar, G. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry*, *12*, 232-241.

Sheehan, D., Lecrubier, Y., Sheehan, K., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. (1998). The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*, 22-33.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*, 3940-3941.

Smits, N., Zitman, F., Cuijpers, P., Den Hollander-Gijsman, M., & Carlier, I. (2012). A proof of principle for using adaptive testing in routine outcome monitoring: The efficiency of the Mood and Anxiety Symptoms Questionnaire–Anhedonic Depression CAT. *BMC Medical Research Methodology*, *12*, 2.

Thompson, N. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*, 778-793.

Thompson, N. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation*, *16*(4), 2.

Ueno, M., & Songmuang, P. (2010). Computerized adaptive testing based on decision tree. In *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 191-193). Los Alamitos, CA: IEEE Computer Society Press.

Van der Linden, W., & Glas, C. (2010). *Elements of adaptive testing*. New York, NY: Springer.

Van Vliet, I., & De Beurs, E. (2007). Het Mini Internationaal Neuropsychiatrisch Interview (MINI). Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV en ICD-10-stoornissen [The Mini International Neuropsychiatric Interview (MINI). A short structured diagnostic psychiatric interview for DSM-IV and ICD-10 disorders]. *Tijdschrift voor Psychiatrie*, *49*, 393-397.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.

Wardenaar, K., van Veen, T., Giltay, E., de Beurs, E., Penninx, B., & Zitman, F. (2010). Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). *Psychiatry Research*, *179*, 101-106.

Ware, J., Jr., Kosinski, M., Bjorner, J., Bayliss, M., Batenhorst, A., Dahlöf, C., . . .Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, *12*, 935-952.

Watson, D., & Clark, L. A. (1991). *The Mood and Anxiety Symptom Questionnaire*. Unpublished manuscript, Department of Psychology, University of Iowa, Iowa City.

Watson, D., Weber, K., Assenheimer, J., Clark, L., Strauss, M., & McCormick, R. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology*, *104*, 3-14.

Weiss, D., & Kingsbury, G. (2005). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375.

World Health Organization. (1993). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. Geneva, Switzerland: Author.

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