STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4311–4327
Published online 14 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2853

Item reduction in a scale for screening

Xinhua Liu*,† and Zhezhen Jin

Department of Biostatistics, Columbia University, New York, NY 10032, U.S.A.

SUMMARY

This paper presents a non-parametric approach for the selection of items in a scale for screening, with the score defined as the sum of item response indicators. Without specifying parametric models for binary classification probabilities, the proposed item selection method evaluates the change in classification accuracy due to adding or deleting one item for a scale with k items. It first removes least useful items from the scale and then uses a forward stepwise selection procedure to the remaining items to identify a subset of items for a reduced scale. The reduced scale usually retains or improves classification accuracy compared to the full scale. The variation in items selected can be assessed with bootstrap samples. In a simulation study, the proposed procedure shows a fairly good finite sample performance. The method is illustrated with a data set on patients with and without high risk of developing Alzheimer's disease who were administered a 40-item test of olfactory function. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: item selection; reduced scale; classification accuracy

1. INTRODUCTION

In mental health studies, an underlying trait of interest is often not directly observable and scales are commonly used to measure it. The scales are composed of a set of items that are correlated to the latent trait. The scale score, often defined as a sum of the response indicators of the items, is usually used as a surrogate measure of the latent trait. One rationale for the use of scale score is that the sum of item responses is a sufficient statistic for the underlying latent trait if the items and the latent trait satisfy the Rasch model [1, 2] for uni-dimensional scale. The scale score can also be used to assess the underlying latent trait when the items satisfy the monotone homogeneity model in non-parametric item response theory (IRT) [3].

*Correspondence to: Xinhua Liu, Department of Biostatistics, Columbia University, New York, NY 10032, U.S.A.
†E-mail: XL26@columbia.edu

Contract/grant sponsor: NSF; contract/grant number: DMS-0134431
Contract/grant sponsor: National Institute on Aging; contract/grant number: R01-AG17761

Received 23 November 2005
Accepted 12 January 2007

A uni-dimensional scale is particularly useful in screening for individuals at risk of a certain illness if it measures the underlying latent trait of the illness. In the ideal case, all items in a scale are consistent, highly correlated with the latent variable, but less correlated to each other. The items should also have high classification ability, e.g. high sensitivity and specificity or predictive values [4]. It is possible, however, that a scale is composed of some 'noisy' items that are not useful in classification. To improve the utility of a scale for specific screening purpose, it is necessary to evaluate all the items in the scale and identify those useful for classification. Excluding uninformative items from the scale will result in a reduced scale that may maintain or even improve the classification accuracy of the full scale. Composed of fewer items, the reduced scale will be more useful in screening for the illness not only because it is easier and quicker to administer than the full scale but also because it leads to cost reductions.

Studies have shown that the development of Alzheimer's disease (AD) is associated with olfactory identification deficits [5, 6]. The deficits are usually measured by the standardized University of Pennsylvania Smell Identification Test (UPSIT) [7]. This self-administered 40-item test (see Appendix) is a uni-dimensional scale with binary responses on the odour items. The test score is the total number of odours correctly recognized. It needs about 30 minutes to complete the test. To improve the clinical utility of UPSIT in screening patients at high risk of AD, clinicians are interested in selecting items from the full scale so that the reduced scale has similar or improved classification ability compared to the full scale. For screening purpose, using number of correctly recognized odours for the scale score is more convenient and meaningful than using weighted sum of responses in odour identification. Particularly, in classifying patients with high *versus* low risk of AD, it makes no sense to use weights with opposite signs for item responses in odour identification. This implies that a reduced scale for screening should be uni-dimensional with equally weighted items. The example demonstrates specific requirements for item reduction in a scale for screening.

In item analysis, IRT has been studied extensively in the fields of education and psychology [2]. In IRT, item response models have parameters characterizing the relationship between the particular items and a latent trait, assuming that item responses are mutually independent conditional on the latent trait. The parameters of interest may include a location and slope for each item. Reliable estimates of the model parameters usually require a very large sample size. Differences in the item-specific parameters between two classes of disease/disorder status may indicate differential item functioning (DIF). The method is useful in item evaluation but not useful in item selection for classification purpose because unequal mean scores distinguishing two classes may not imply DIF, and DIF can occur even when within-class mean scores are similar [8].

Classification and regression tree analysis [9] are often used in selecting variables for classification. With preset criteria, items useful for classification can be selected for splits in the tree structure. The method also requires a very large sample size, while the number of selected items tends to be small, due to possible multiple uses of some items in the sub-trees. A tree structured scale is more complicated to use than the sum of item response indicators. Ignoring tree structures, using sum of responses on the few selected items cannot have good classification ability.

In classifying two classes, when the item responses are binary variables, logistic regression analysis [10] is more proper than discrimination analysis [11], which assumes normality for quantitative classifiers [12]. Logistic regression, using logit link between probability of disease class and a linear combination of predictors, can be applied to data from prospective as well as case–control studies, and under some conditions [13] can perform well even when the logistic model is not correct. To identify the items jointly predictive for disease/disorder status, various

DOI: 10.1002/sim

variable selection procedures in logistic regression analysis are available in most statistical software packages, including forward, backward and stepwise selection procedures based on score test and Wald test for hypotheses on the regression parameters [14]. The subset of items selected from the full scale will depend on the model, the selection procedure and the selection criteria. It is known that none of the selection procedures can guarantee unidirectional estimates for the coefficients of the response indicators of the selected items, a necessary condition for the uni-dimensional scale with equally weighted items.

To identify the items distinguishing two classes in the same direction for a reduced scale, one may use one-sided tests for each item, with proper controls of family wise error rate (FWER) or false discovery rate (FDR) for the multiple tests [15]. The control may be based on either re-sampling or raw p-values obtained from correlated one-sided tests [16]. Note that larger between-item correlations may cause redundancy among the items selected by the multiple tests based procedure. A selection procedure in logistic regression analysis may help remove redundant items to obtain a final subset of the items. This two-stage approach may greatly increase the likelihood, while still cannot guarantee for the selected items to have unidirectional estimates of the coefficients [17].

Recently, Pepe *et al.* [18] showed that fitting a logistic regression model to data could yield poor classification performance. Hastie *et al.* [19] also noted that it is hard to verify regression models, especially in high dimensions. In general, regression model-based item selection requires correct model specification. To satisfy the necessary condition for uni-dimensional scale with equally weighted items, model parameters have to be constrained in the same direction. This may greatly complicate parametric model-based selection procedures.

In this paper, we propose a simple non-parametric approach for selecting items for a parsimonious uni-dimensional scale with equally weighted items useful for screening specific illness. The reduced scale has fewer items than the full scale but retains or even improves classification ability.

The criteria useful to evaluate the classification ability of a scale include prediction or classification error [19], receiver operating characteristic (ROC) curve and the area under the ROC curve [4, 20]. The ROC curve shows how sensitivities change either with specificities or with false-positive proportions (1-specificities) for all possible cutoff scores, where pairs of sensitivity and specificity are defined at the same cut-off point. The area under the ROC curve (AUC) is a summary of the ROC curve. It represents the probability that the measure from a subject with the illness indicates a greater suspicion than that from a subject without the illness [21]. In this paper, the classification accuracy is defined similarly to AUC, as the probability that the score from a subject randomly selected from one class is less than that from the subject randomly selected from the counterpart class. Based on evaluation of the change in classification accuracy due to inclusion or exclusion of an item, we select items for a reduced uni-dimensional scale. In the next section, we present the proposed method. Then we evaluate finite sample performance of the proposed procedure with a simulation study in Section 3. In Section 4, we illustrate the method using data from a study of the 40-item scale (UPSIT) administered to test olfactory functioning in patients with and without high risk of AD in a follow-up study [17]. Section 5 presents some discussions.

2. METHOD

Suppose that a full scale has m items in the set W_m and response on each item is binary, i.e. $X_t \in \{0, 1\}$ for item t, $t = 1, \ldots, m$. The score on the full scale, defined as $S(W_m) = \sum_{t=1}^{m} X_t$, takes integer values between zero and m. The score on a reduced scale $S(W_k)$, defined on k items

DOI: 10.1002/sim

in $W_k \subset W_m$, $1 \leqslant k < m$, can be viewed as a weighted sum of responses on all items in W_m,

$$S(W_k) = \sum_{t=1}^{m} I(\text{item}_t \in W_k) X_t$$

where $I(\text{item}_t \in W_k)$ takes value 1 if item_t in the reduced scale, and 0 otherwise.

Suppose that subjects with the illness ($D = 1$) tend to have higher score than the subjects without the illness ($D = 0$). For a scale with k items in W_k, let $S^D(W_k)$ be the score for subject in class D, $1 \leqslant k \leqslant m$, $D = 0, 1$; we define classification accuracy as

$$\text{CA}(W_k) = P(S^0(W_k) < S^1(W_k))$$

with values ranging between zero and one. Note that $\text{CA}(W_k)$ is part of the AUC for a score with k items in W_k defined by using linear interpolation between adjacent points [20],

$$\text{AUC}(W_k) = \text{CA}(W_k) + P(S^0(W_k) = S^1(W_k))/2$$

The ROC curve is defined on the set of points $\{(\text{FPP}(c), \text{TPP}(c)); c = 0, 1, \ldots, k\}$, with $\text{TPP}(c) = P(S^1(W_k) \geqslant c)$ as true positive proportion or sensitivity, and $\text{FPP}(c) = P(S^0(W_k) \geqslant c)$ as false-positive proportion or 1-specificity. Therefore, $\text{CA}(W_k)$ retains the invariance property of $\text{AUC}(W_k)$, i.e. the classification accuracy only depends on the ranks of $S(W_k)$s in the sense that the untransformed score $S(W_k)$ and any transformed score $g(S(W_k))$ with function g that reserves the rank of $S(W_k)$ would lead to the same $\text{CA}(W_k)$.

2.1. The change in classification accuracy

Let X_{it}^D be the response indicator of item t for subject i in class D, $j = 1, \ldots, m$; $i = 1, \ldots, n_D$; $D = 0, 1$. Then the score for the scale with items in $W_k \subset W_m$, $1 \leqslant k < m$, from the ith subject in class D, is

$$S_i^D(W_k) = \sum_{t=1}^{m} X_{it}^D I(\text{item}_t \in W_k)$$

We may estimate classification accuracy for the scale with the k items in W_k by

$$A(W_k) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(S_i^0(W_k) < S_j^1(W_k))$$

The estimator has a form similar to the area under empirical ROC curve [22]. As a U-statistic, $A(W_k)$ is easy to calculate and its statistical properties are easy to establish. Obviously, $E[A(W_k)] = \text{CA}(W_k)$.

Note that the test score on the scale with k items in W_k for subject i in class D can be expressed as $S_i^D(W_k) = S_i^D(W_k \setminus \{\text{item}_h\}) + X_{ih}^D$, where $\text{item}_h \in W_k$, $1 < k \leqslant m$. This relationship is useful to calculate the changes in estimated classification accuracy, either

$$\Delta_k(-X_h | W_k) = A(W_k) - A(W_k \setminus \{\text{item}_h\})$$

due to excluding item_h from the item set W_k, or

$$\Delta_k(+X_h | W_{k-1}) = A(W_{k-1} \cup \{\text{item}_h\}) - A(W_{k-1})$$

due to adding item_h into W_{k-1} for a new set W_k.

DOI: 10.1002/sim

Let $d_{ij}(k) = S_j^1(W_k) - S_i^0(W_k)$ and $z_{ij}(h) = X_{jh}^1 - X_{ih}^0$, then

$$\Delta_k(-X_h|W_k) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}[I\{z_{ij}(h)>0,\ d_{ij}(k)=1\} - I\{z_{ij}(h)<0,\ d_{ij}(k)=0\}]$$

$$= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} U_{-h}(i, j|W_k)$$

and

$$\Delta_k(+X_h|W_{k-1}) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}[I\{z_{ij}(h)>0,\ d_{ij}(k-1)=0\} - I\{z_{ij}(h)<0,\ d_{ij}(k-1)=1\}]$$

$$= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} U_{+h}(i, j|W_{k-1})$$

In summary, we can write the change $\Delta = A(W_k) - A(W_{k-1})$ for $W_{k-1} \subset W_k$ in the form of U-statistic that

$$\Delta = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} U_{ij}$$

where U_{ij} can be either $U_{-h}(i, j|W_k)$ or $U_{+h}(i, j|W_{k-1})$. The change in estimated classification accuracy has mean $\mu = E(\Delta) = E(U_{ij})$ and variance

$$\mathrm{se}^2(\Delta) = \frac{(n_0-1)\lambda_{01} + (n_1-1)\lambda_{10} + \lambda_{11} - (n_0+n_1-1)\mu^2}{n_0 n_1}$$

where $\lambda_{01} = E(U_{ij}U_{ik})$, $j \neq k$; $\lambda_{10} = E(U_{ij}U_{hj})$, $i \neq h$; and $\lambda_{11} = E(U_{ij}^2)$ for $i, h = 1, \ldots, n_0$; $j, k = 1, \ldots, n_1$. By the Central Limit Theory on U-statistic, as $n_D \to \infty$, $D = 0, 1$, we have

$$\sqrt{n_0 n_1}\ (A(W_k) - A(W_{k-1})) \to \mathrm{N}(\mu_k, \sigma_k^2)$$

where $W_{k-1} \subset W_k$ and $\mu_k = \mathrm{CA}(W_k) - \mathrm{CA}(W_{k-1})$ is the change in classification accuracy due to inclusion or exclusion one item from a scale with k items. It is easy to see that $\mathrm{se}(\Delta)$ can be estimated by plugging in the empirical counterparts of λ_{01}, λ_{10} and λ_{11}. We denote the estimator as $\hat{\mathrm{se}}(\Delta)$.

Under the null hypothesis $H_0: \mu_k = 0$ with $W_{k-1} \subset W_k$, the test statistic,

$$\mathrm{TS} = \frac{A(W_k) - A(W_{k-1})}{\hat{\mathrm{se}}(A(W_k) - A(W_{k-1}))}$$

will have approximate N(0, 1) distribution. We propose to use it as a basis for item selection along with a prespecified threshold value γ. The relevant hypotheses are $H_0: \mu_k \leqslant 0$ *versus* $H_1: \mu_k > 0$.

Particularly, we will use a prespecified threshold value γ_0 and the test statistic

$$\mathrm{TS}(-X_j|W_k) = \frac{\Delta_k(-X_j|W_k)}{\hat{\mathrm{se}}(\Delta_k(-X_j|W_k))}$$

DOI: 10.1002/sim

to decide whether or not to remove item$_j$ from W_k. When $\mathrm{TS}(-X_j|W_k)<\gamma_0$, we will exclude item$_j$ from W_k. Similarly, we will use a prespecified threshold value γ_1 and the test statistic

$$\mathrm{TS}(+X_h|W_{k-1})=\frac{\Delta_k(+X_h|W_{k-1})}{\hat{\mathrm{se}}(\Delta_k(+X_h|W_{k-1}))}$$

to decide whether or not to add item$_h \in W_m \backslash W_{k-1}$ into W_{k-1} for a new set W_k. We will have item set $W_k=\{\text{item}_h\}\cup W_{k-1}$ when $\mathrm{TS}(+X_h|W_{k-1})\geqslant\gamma_1$.

2.2. Item selection procedure

It is obvious that item$_j$ in W_k for a scale with score $S(W_k)$ is least useful for classification if excluding it from W_k leads to no change or an increase in the estimated classification accuracy or $\Delta_k(-X_j|W_k)\leqslant 0$. Consequently, we first identify the least useful items from the full scale, if any, and then apply hypothesis test-based stepwise selection procedure to the remaining items.

Starting with the item set W_m for the full scale, we will identify the least useful item, if any, and remove it. For $1<k\leqslant m$, we will exclude item$_{k_0}$ from W_k if the corresponding change in estimated classification accuracy $\Delta_k(-X_{k_0}|W_k)\leqslant 0$, where

$$\Delta_k(-X_{k_0}|W_k)=\min_{\text{item}_j\in W_k}\{A(W_k)-A(W_k\backslash\{\text{item}_j\})\}$$

The deletion process will stop when no more items can be removed. The resulting item set is denoted as W_J, $1<J<m$. This process will produce a sequence of subsets $\{W_k;\ J<k\leqslant m\}$ such that $W_J\subset\cdots\subset W_m$, with a sequence of estimated classification accuracies $\{A(W_k);\ J<k\leqslant m\}$ satisfying $A(W_J)\geqslant\cdots\geqslant A(W_m)$.

Although the item set W_J has less items, it might still have some unstable items that make little contributions to the classification accuracy. It is important to identify relatively stable items in W_J to form a further reduced scale without sacrificing much in classification accuracy. This can be accomplished by the following hypothesis test-based selection procedure along with prespecified positive threshold values γ_0 and γ_1 ($\gamma_0\leqslant\gamma_1$):

(i) Identify the item in W_J that has the largest value of the test statistic for classification accuracy. Let Q_1 denote the resulting singleton item set. Specifically, the initial item set $Q_1=\{\text{item}_h\}\subset W_J$ satisfying

$$\mathrm{TS}(A(Q_1))=\max_{\text{item}_j\in W_J}\mathrm{TS}(A(\{\text{item}_j\}))\geqslant\gamma_1$$

where $\mathrm{TS}(A(\{\text{item}_j\}))=A(\{\text{item}_j\})/\hat{\mathrm{se}}(A(\{\text{item}_j\}))$ with

$$A(\{\text{item}_j\})=\sum_{i=1}^{n_0}\sum_{h=1}^{n_1}\frac{I(X_{ij}^0<X_{hj}^1)}{n_0n_1}=\sum_{i=1}^{n_0}\frac{I(X_{ij}^0=0)}{n_0}\sum_{h=1}^{n_1}\frac{I(X_{hj}^1=1)}{n_1}=\hat{\mathrm{Sp}}(j)\hat{\mathrm{Se}}(j)$$

and

$$\hat{\mathrm{se}}^2(A(\{\text{item}_j\}))=\frac{\hat{\mathrm{Se}}(j)\hat{\mathrm{Sp}}(j)}{n_0n_1}[1+(n_0-1)\hat{\mathrm{Sp}}(j)+(n_1-1)\hat{\mathrm{Se}}(j)$$

$$-(n_0+n_1-1)\hat{\mathrm{Sp}}(j)\hat{\mathrm{Se}}(j)]$$

DOI: 10.1002/sim

Here $\hat{\mathrm{Sp}}(j)$ and $\hat{\mathrm{Se}}(j)$ are the estimators of specificity $\mathrm{Sp}(j)=P(X_j^0=0)$ and sensitivity $\mathrm{Se}(j)=P(X_j^1=1)$ for item$_j$, respectively.

(ii) For $1<k\leqslant J$, identify the item item$_{k_0}$ that has the largest value of the test statistic for H_0: $\mu_k\leqslant 0$ *versus* H_1: $\mu_k>0$ from the item set $W_J\setminus Q_{k-1}$. Let $Q_k=Q_{k-1}\cup\{\text{item}_{k_0}\}$ if

$$\mathrm{TS}(+X_{k_0}|Q_{k-1})=\max_{\text{item}_j\in W_J\setminus Q_{k-1}}\mathrm{TS}(+X_j|Q_{k-1})\geqslant\gamma_1$$

(iii) Identify unstable items in Q_k that their removal leads to little loss or even an improvement in estimated classification accuracy. Specifically, item$_h$ is excluded from Q_k, $1<k\leqslant J$, if

$$\mathrm{TS}(-X_h|Q_k)=\min_{\text{item}_j\in Q_k}\mathrm{TS}(-X_j|Q_k)<\gamma_0$$

The exclusion process will stop if no more items can be removed.

(iv) Repeat steps (ii) and (iii) until no more items could be added and removed or stop the process if an item that has been removed tends to be added in again.

The final item set is denoted as Q_H. We will use the items in Q_H for a reduced scale.

2.3. Evaluation of variability in selected items

In studies of rare diseases and uncommon conditions, the sample used to select items in a scale for screening may not be very large. The variation of selected items can be assessed by bootstrap approach [23] with the proposed procedure. Bootstrap samples can be obtained by sampling with replacement from the original study sample, where the sampling unit is the study subject with a cluster of observed responses on the items in the full scale. The empirical distributions of the number of selected items, the estimated classification accuracy for the full scale, the reduced scale, and the difference in estimated classification accuracies between the full and reduced scales can be used for inference. Moreover, the selection frequency of each item in a number of bootstrap samples (say 1000) provides an empirical estimate of how often an item is selected. The spectrum of item selection may help identify the most frequently selected items.

3. A SIMULATION STUDY

To examine the finite sample performance of the selection procedure with different thresholds, we conducted a simulation study for a hypothetical uni-dimensional scale with 13 items, among which six items are useful for distinguishing two classes. The data were generated from logistic models in which logit function links item response probabilities to an underlying latent trait Z,

$$\text{logit}\,P(X_t=1|D,Z)=\alpha_t^D+\beta_t^D Z,\quad t=1,\ldots,13,\quad D=0,1$$

where α_t^D is the location parameter and β_t^D is the slope parameter. The β_t^D reflects the degree of association between item responses and the latent trait. The larger the β_t^D is, the stronger the association is. Table I lists preset values for α_t^D and β_t^D used in data generation. We preset six items, $\{8,\ldots,13\}$ with different parameters by class. The two classes have equal sample size, $N=50$, 100, and 200. In each case, we generated 1000 data sets. In each data set, we first

DOI: 10.1002/sim

Table I. Parameters in logistic models for data generation.

Item t	1	2	3	4	5	6	7	8	9	10	11	12	13
α_t^0	−2	−1	−1	−1	−0.5	−0.5	−0.5	−2	−2	−2	−2	−2	−2
β_t^0	1	1	2	3	1	2	3	1	1	1	1	1	1
α_t^1	−2	−1	−1	−1	−0.5	−0.5	−0.5	−1	−1	−1	−0.5	−0.5	−0.5
β_t^1	1	1	2	3	1	2	3	1	2	3	1	2	3

Table II. Proposed method: performance of selected scales.

Criteria (γ_0, γ_1)	Scale size H mean (SD)	$A(W_H)$ mean (SD)	Δ (per cent) mean (SD)	$\Delta>0$ (per cent)
$N=50$				
Full scale	13	0.5485 (0.0577)		
$W=\{8,\ldots,13\}$	6	0.6043 (0.0583)	10.46 (6.45)	96.4
I: (0.52, 0.5244)	4.75 (1.04)	0.6148 (0.0566)	12.41 (5.99)	99.5
II: (0.84, 0.8416)	4.04 (0.85)	0.6054 (0.0583)	10.65 (6.32)	97.5
III: (1.036, 1.0364)	3.86 (0.79)	0.5977 (0.0588)	9.24 (6.65)	93.0
IV: (1.28, 1.2816)	3.19 (0.67)	0.5849 (0.0608)	6.90 (7.07)	84.4
$N=100$				
Full scale	13	0.5476 (0.0425)		
$W=\{8,\ldots,13\}$	6	0.6025 (0.0419)	10.16 (4.20)	99.4
I: (0.52, 0.5244)	5.37 (0.89)	0.6067 (0.0413)	10.93 (3.96)	100
II: (0.84, 0.8416)	4.85 (0.82)	0.6024 (0.0418)	10.15 (4.15)	99.8
III:(1.036, 1.0364)	4.52 (0.78)	0.5984 (0.0424)	9.40 (4.33)	99.1
IV: (1.28, 1.2816)	4.13 (0.75)	0.5922 (0.0426)	8.27 (4.45)	97.8
$N=200$				
Full scale	13	0.5485 (0.0309)		
$W=\{8,\ldots,13\}$	6	0.6044 (0.0311)	10.28 (3.19)	100
I: (0.52, 0.5244)	5.72 (0.73)	0.6059 (0.0311)	10.54 (3.00)	100
II: (0.84, 0.8416)	5.38 (0.71)	0.6041 (0.0312)	10.21 (3.06)	100
III: (1.036, 1.0364)	5.20 (0.71)	0.6027 (0.0314)	9.96 (3.08)	100
IV: (1.28, 1.2816)	4.91 (0.73)	0.5999 (0.0329)	9.44 (3.19)	100

Note: $A(W_H)$: estimated classification accuracy of selected scale. $\Delta=(A(W_H)-A(W_{13}))/A(W_{13})\times 100$ per cent: per cent improvement.

generated $2N$ independent random numbers from a standard normal distribution for latent trait Z. Then for each value of Z, we generated 13 independent binary responses with probabilities specified by the logistic models. We apply the proposed procedure with four sets of threshold values $(\gamma_0, \gamma_1)=(0.52, 0.5244)$, $(0.84, 0.8416)$, $(1.036, 1.0364)$, $(1.28, 1.2816)$ according to 70th, 80th, 85th, and 90th percentiles of standard normal distribution, respectively.

Table II showed that the estimated classification accuracies $A(W_{13})$ for the full scale, and $A(W_6)$ for the preset six-item scale with $W_6=\{X8,\ldots,X13\}$, all had consistent mean regardless of sample size. The standard deviations of these estimated quantities, however, decreased with increasing

DOI: 10.1002/sim

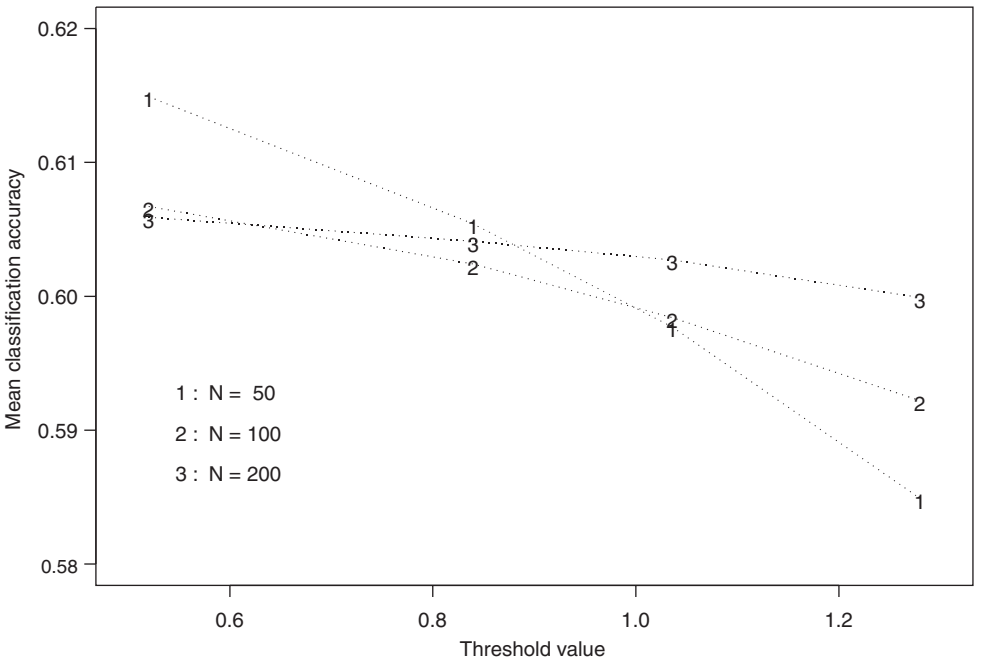

Figure 1. Mean classification accuracy of reduced scales.

sample size. It is not surprising that the improvement in estimated classification accuracy with the preset scale was not always positive when class size is not large.

As expected, with a given data set, the items selected with higher threshold values were in a subset of the items selected with lower threshold values. Consequently, increasing threshold values resulted in fewer selected items and lower classification accuracies. The trend was clear in the cases with class size of 50 while became less apparent as class size increased to 200 (Figure 1). Among the four sets of threshold values, $(\gamma_0, \gamma_1) = (0.84, 0.8416)$ yielded results similar to that of the preset six-item scale.

Table III listed the frequencies of items being selected based on different selection criteria. It is interesting to note that in all the cases, the most frequently selected items were the six preset items useful for classification. The frequencies of correctly selected items increased with sample size while decreased with increasing threshold values. In contrast, the frequencies of incorrectly selected items decreased with sample size or threshold values increased.

As a reviewer suggested, we compared performance of proposed method with that of logistic model-based forward and backward selection procedures for item selection using the same 1000 generated data sets with class size of 100.

The forward selection procedure computed the score test statistic on null hypothesis of zero coefficient for each item not in logistic model and identified the largest of these statistics. If it was significant at the preset level, then the corresponding item was added into the model. The process was repeated until none of the remaining effects met the specified criterion. The backward selection procedure started with fitting logistic model with all items. Wald test on null hypothesis of zero coefficient for each item was examined. The least significant item that did not meet the preset

DOI: 10.1002/sim

Table III. Proposed method: frequency of items selected in 1000 simulated data sets.

	Item												
	1	2	3	4	5	6	7	8	9	10	11	12	13
$N=50$ (γ_0, γ_1)													
I: (0.52, 0.5244)	102	96	29	12	120	27	15	622	557	584	910	843	830
II: (0.84, 0.8416)	48	55	19	7	61	16	10	504	445	485	856	766	766
III: (1.036, 1.0364)	34	37	10	4	41	11	8	432	377	428	812	729	735
VI: (1.28, 1.2816)	13	17	5	3	29	5	8	337	305	366	743	681	679
$N=100$ (γ_0, γ_1)													
I: (0.52, 0.5244)	74	80	10	1	113	14	6	785	667	717	988	962	951
II: (0.84, 0.8416)	35	38	5	1	72	9	2	681	552	619	976	944	914
III: (1.036, 1.0364)	24	26	5	0	44	7	0	602	491	547	962	917	890
VI: (1.28, 1.2816)	12	14	4	0	22	5	0	499	395	486	946	879	870
$N=200(\gamma_0, \gamma_1)$													
I: (0.52, 0.5244)	51	48	0	0	78	2	0	920	803	824	1000	997	995
II: (0.84, 0.8416)	23	23	0	0	42	1	0	870	707	737	999	994	988
III: (1.036, 1.0364)	11	19	0	0	29	0	0	831	655	685	999	990	977
VI: (1.28, 1.2816)	3	14	0	0	16	0	0	756	570	607	998	982	962

Table IV. Logistic model based selections: performance of selected scales.

Class size $N=100$	Scale H mean (SD)	$A(W_H)$ mean (SD)	Δ (per cent) mean (SD)	per cent ($\Delta>0$)
Full scale	13	0.5476 (0.0425)		
Backward selection				
$\alpha=0.05$	5.82 (1.35)	0.5279 (0.0504)	−3.6538 (4.3790)	20.8
$\alpha=0.10$	6.98 (1.36)	0.5356 (0.0474)	−2.2287 (3.2228)	25.3
$\alpha=0.15$	7.74 (1.39)	0.5386 (0.0462)	−1.6783 (2.7892)	29.5
Forward selection				
$\alpha=0.05$	5.69 (1.39)	0.5260 (0.0507)	−4.0046 (4.5413)	18.8
$\alpha=0.10$	6.87 (1.44)	0.5347 (0.0478)	−2.3974 (3.3585)	23.4
$\alpha=0.15$	7.67 (1.44)	0.5378 (0.0465)	−1.8213 (2.8763)	27.8

Note: $A(W_H)$, estimated classification accuracy; Δ, per cent improvement.

level for staying in the model was removed. The process of model fitting and testing individual effect of items was repeated until no other effect in the model could meet the specified level for removal.

Tables IV presents the descriptive statistics for the number of items selected, the estimated classification accuracy and per cent improvement for the reduced scales with items selected by logistic model-based selection procedures using commonly accepted criteria on significance level of two-sided tests, 0.05, 0.10, 0.15. The two model-based selection procedures yielded similar results. The number of selected items and the estimated classification accuracy of the reduced scales increased with significance level. They seemed to be slightly larger with backward selection than with forward selection procedure. Compared to the full scale, the reduced scales tended to

DOI: 10.1002/sim

Table V. Logistic model based selection: frequency of items selected in 1000 simulated data sets.

	Item number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Backward selection													
$\alpha = 0.05$	111	147	336	529	150	340	495	293	377	550	815	814	859
$\alpha = 0.10$	185	224	428	630	211	427	591	425	489	674	891	885	919
$\alpha = 0.15$	249	284	507	673	261	506	636	524	573	744	928	915	942
Forward selection													
$\alpha = 0.05$	113	149	335	516	147	333	474	279	352	543	807	802	844
$\alpha = 0.10$	184	216	429	609	208	422	580	409	477	653	886	883	917
$\alpha = 0.15$	249	277	504	668	259	501	633	510	559	734	921	910	942

Note: Class size $N = 100$.

have somewhat lower classification accuracy. Table V showed that with the logistic model-based procedures, not all important items were more frequently chosen compared to the unimportant ones.

In summary, the proposed procedure showed a fairly good finite sample performance in the simulation study. Empirically, the threshold values could be in the range of 0.6–1.3. To reduce the possibility of excluding useful items, one may choose several threshold values and examine the corresponding results. Nevertheless, increasing class size may reduce the impact of threshold values and give more stable result.

4. APPLICATION

To illustrate the proposed method, we use the olfaction test data collected from 127 patients with mild cognitive impairment (MCI) who were at risk to develop AD [17]. The patients were administered UPSIT, a 40-item olfactory test, at baseline, and followed for at least two years. There were 31 patients who met criteria of AD diagnosis within two years after baseline evaluation, being considered in the group with high risk of developing AD ($D = 1$). The low-risk group ($D = 0$) had 96 patients who did not develop AD in the two years of follow-up.

The purpose of the analysis is to identify the items in the 40-item olfactory test for a reduced scale that may efficiently classify the two groups of MCI patients. The test score of UPSIT is the sum of odours correctly recognized by a patient, with low score indicating poor olfactory functioning. For our purposes, we used a scale score defined as the number of odours incorrectly identified such that a high score indicates poor olfactory functioning.

In this sample, Cronbach's coefficient alpha estimate [24] for UPSIT items was 0.8843, indicating a good consistency of items in relation to the underlying trait of olfactory functioning. Figures 2 and 3 show a wide range in the proportion of incorrect odour recognition (9.45–64.57 per cent), in the item sensitivity (16.13–74.19 per cent), in the item specificity (37.5–93.75 per cent), and in the item classification accuracy (0.1378–0.4476). Most items had high specificity with low sensitivity.

We first used the popular logistic model-based selection procedures for item selection with commonly used inclusion and exclusion criteria $\alpha_{\text{in}} = \alpha_{\text{out}} = \alpha = 0.10$. The backward selection

DOI: 10.1002/sim

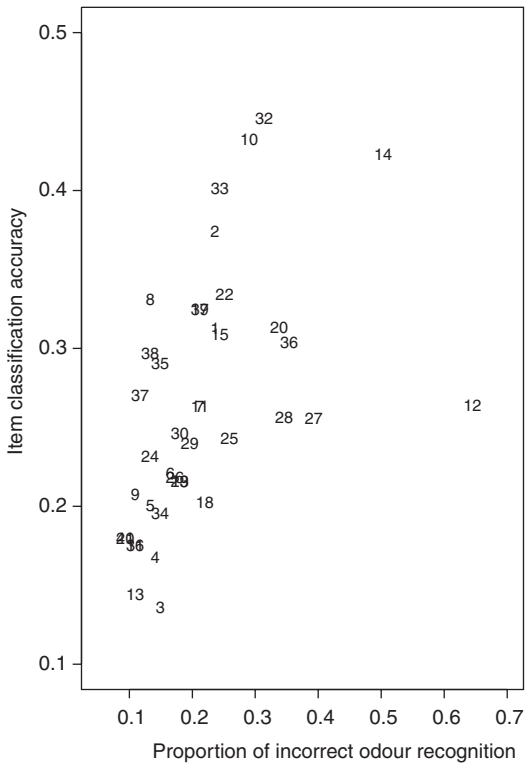

Figure 2. Item classification accuracy and proportion of response.

procedure selected eight items (X4, X5, X8, X10, X14, X21, X32, X33) with estimated coefficients for X4 and X5 in the opposite direction of the other six estimated coefficients. The forward and stepwise selection procedures selected six items (X8, X10, X16, X32, X33, X38) with estimated coefficient for X16 in the opposite direction of the other five estimated coefficients. The result implied that the highest likelihood for being in the high-risk class was not given to those who could not correctly recognize any items in the selected subset of the odours. Obviously, this was not biologically meaningful.

Alternatively, we applied logistic model-based selection procedures only to the items with potential to distinguish two classes in the same direction. By one-sided Fisher exact test, there were 18 pre-selected items with raw p-values below 0.05, while using the resampling technique [13] the multiple test adjusted p-values were less than 0.50. With $\alpha = 0.10$ for the backward, forward, and stepwise logistic regression selection procedures, we obtained the same six-item subset $W_6 = \{X8,\ X10,\ X21,\ X32,\ X33,\ X38\}$ with estimated coefficients in the same direction. The estimated classification accuracy of the six-item scale was 0.7984, exceeding the 0.7678 ($\hat{\text{se}} = 0.0450$) of the 40-item full scale. This result suggested that the six odours could be used for a reduced uni-dimensional scale to screen patients at high risk of AD, with number of incorrectly identified odours as test score. It is interesting to note that the result was invariant when the threshold value was changed to $\alpha = 0.15$.

DOI: 10.1002/sim

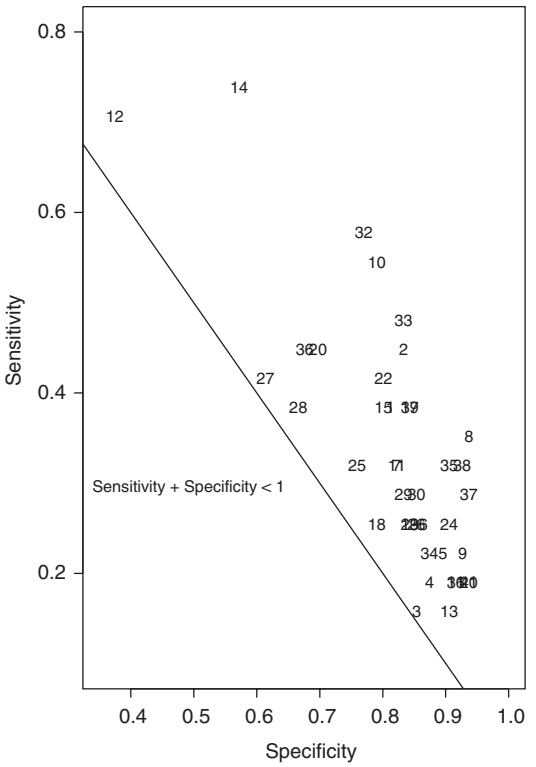

Figure 3. Item sensitivity and specificity.

In contrast, the proposed method identified 20 items as least useful items to be excluded from the full scale. With threshold values $(\gamma_0, \gamma_1) = (0.84,\ 0.8416)$, the stepwise selection procedure suggested inclusion of $H = 7$ items for a reduced scale that $W_7 = \{X8,\ X10,\ X21,\ X32,\ X33,\ X37,\ X38\}$. The estimated classification accuracy $A(W_7)$ was 0.8165, also exceeding the 0.7678 of the full scale. The result was unchanged when the threshold values were increased to $(\gamma_0, \gamma_1) = (1.036, 1.0364)$.

To evaluate the variation in item selections, we applied the proposed selection procedure to 1000 bootstrap samples obtained from the original data set ($n = 127$). Table VI shows that the median number of selected items in the bootstrap samples was six, and the estimated classification accuracy with the reduced scales, $A(W_H)$, was consistently larger than that produced using the full scale. Specifically, in 99.7 per cent of the bootstrap samples $A(W_H)$ was larger than $A(W_{40})$. This suggested that the 40-item scale could be greatly reduced for the binary classification. Figure 4 shows the item selection spectrum of the bootstrap samples. Thirty-two items had been selected with low frequencies ($\leqslant$23.1 per cent of the bootstrap samples), items $X14$ and $X21$ were selected in 47.4 and 41.7 per cent of the bootstrap samples, respectively. The most frequently selected six items, selected in 51.5–71.8 per cent of the bootstrap samples, were also in W_7 selected by the proposed procedure using the original data set. It seemed that item X14 might be also important. Nevertheless, to obtain a confirmative result, we need to increase sample size, especially in the high-risk class.

DOI: 10.1002/sim

Table VI. Performance of scales selected in 1000 bootstrap samples.

Statistics	Mean (SD)	Median (range)
Reduced scale size H	6.06 (1.29)	6 (3, 10)
Reduced scale: $A(W_H)$	0.8364 (0.0434)	0.8387 (0.6996, 0.9721)
Full scale: $A(W_{40})$	0.7702 (0.0429)	0.7717 (0.6182, 0.9273)
$A(W_H) - A(W_{40})$	0.0671 (0.0361)	0.0675 (−0.0521, 0.2070)
$\frac{A(W_H)-A(W_{40})}{A(W_{40})} \times 100$ per cent	10.69 (4.84)	10.55 (−3.05, 31.97)

Note: Selection criteria $(\gamma_0,\ \gamma_1) = (0.84, 0.8416)$.

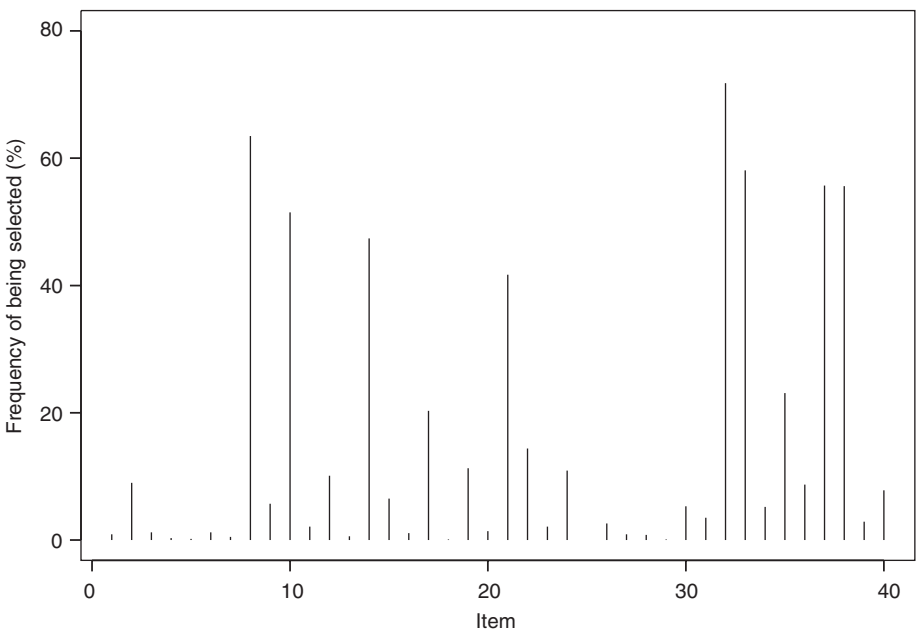

Figure 4. Proportion of items selected in 1000 bootstrap samples.

In this application, item set W_6 obtained by 2-stage selection, in which logistic model-based selection procedures was applied to the items pre-selected with one-sided Fisher exact test, happened to be included in W_7 selected by the proposed method. The parametric model-based selection approaches, however, had different basis from the proposed non-parametric method. Based on evaluation of classification accuracy in the selection process, our approach is more relevant to the goal of selecting items to retain or improve classification accuracy for a uni-dimensional scale with equally weighted items.

DOI: 10.1002/sim

5. DISCUSSION

We have proposed a non-parametric method for item reduction in a uni-dimensional scale for screening, based on evaluation of classification accuracy. Because the classification accuracy is invariant to the rank reserved transformations on the scores, the validity of the proposed item selection procedure depends only on the assumption that

$$P(D=1|X_1,\ldots,X_m)=P\left(D=1\left|\sum_{t=1}^{m}X_t I(\text{item}_t\in W_H)\right.\right)$$

This assumption does not require specifying the relationship between the probability $P(D=1|X_1,\ldots,X_m)$ and the item response indicators $X_1,\ldots,X_m$ explicitly. Consequently, the selection procedure is robust.

Note that examining all the possible item combinations in a search for the set of items that will maximize estimated classification accuracy can be time consuming, especially when the number of items in the full scale is large. Besides, the selected item set may include unstable items that make little contribution to the estimated classification accuracy. In contrast, the proposed method based on test statistics for the hypotheses on the change in classification accuracy could quickly select the items that have good classification ability along with relative stability.

The proposed selection procedure begins with removing the least useful items from the scale and then applies stepwise selection to the remaining items. To decide whether or not to include an item for a reduced scale, the proposed stepwise selection requires pre-specified values for thresholds γ_0 and γ_1. Obviously, smaller threshold values may lead to the selection of more items, of which some could be unstable. On the other hand, higher threshold values may lead to qualifying fewer items for a reduced scale, resulting in a smaller estimate of classification accuracy. An upper bound on the thresholds will be necessary for the estimated classification accuracy of a reduced scale to exceeding that of the full scale. Based on a simulation study, we would empirically recommend using few threshold values between 0.6 and 1.3. Meanwhile, examination of item selection frequencies through bootstrap samples may help assess the variation in the item selection. The most frequently selected items could be used for reduced scale. Another important message from the simulation study is that large sample size in both classes may reduce the impact of threshold values on selection result.

The method is applicable when items in a scale have $K(>2)$ response levels in that $K-1$ binary indicators can be produced; for example, $I(X=j)$ for $j=2,\ldots,K$. If the K response levels are ordinal, we may use the binary indicator $I(X\leqslant j)$ for $j=1,\ldots,K-1$.

Similar to logistic regression analysis, the proposed method can accommodate data from both prospective and case–control studies [25]. To improve the usefulness of a reduced scale for screening, we would prefer a large sample from a prospective study. As item selection results always depend on study samples, the reduced scale is only applicable to screen populations comparable to the study population. Therefore, it is important, though it will be challenging, to further develop methods that may combine data from multiple resources to efficiently identify items for a reduced scale, which will be relatively invariant across populations.

APPENDIX A

Description of the University of Pennsylvania Smell Identification Test (UPSIT). The self-administered 40-item scratch-and-sniff multiple choice odour identification test consists of four

DOI: 10.1002/sim

booklets containing 10 odorant apiece, with one odorant per page. The stimuli are released by scratching each strip with a pencil tip in a standardized manner. Above each odorant strip is a multiple-choice question with four alternative responses for each item. Subjects are instructed to scratch the label, then sniff the label and choose a response category closest to the smell that they experienced.

Item, odorant	Item, odorant	Item, odorant	Item, idorant
X01 pizza	X11 onion	X21 lilac	X31 paint thinner
X02 bubble gum	X12 fruit punch	X22 turpentine	X32 grass
X03 menthol	X13 licorice	X23 peach	X33 smoke
X04 cherry	X14 cheddar cheese	X24 root beer	X34 pine
X05 motor oil	X15 cinnamon	X25 dill pickle	X35 grape
X06 mint	X16 gasoline	X26 pineapple	X36 lemon
X07 banana	X17 strawberry	X27 lime	X37 soap
X08 clove	X18 cedar	X28 orange	X38 natural gas
X09 leather	X19 chocolate	X29 wintergreen	X39 rose
X10 coconut	X20 gingerbread	X30 watermelon	X40 peanut

ACKNOWLEDGEMENTS

This research was partially supported by the NSF Career award DMS-0134431 to Zhezhen Jin. We thank Drs Devanand and Tabert for providing us the data from their prospective study at the Memory Disorders Center of the New York State Psychiatric Institute and Columbia University, funded by National Institute on Aging grant R01-AG17761. We also thank the reviewers for their helpful comments and thank Dr Zeger for his valuable suggestions.

REFERENCES

1. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, MA, 1968.
2. Van Der Linden WJ, Hambleton RK. *Handbook of Modern Item Response Theory*. Springer: New York, 1997.
3. Sijtsma K, Molenaar IW. *Introduction to Nonparametric Item Response Theory*. Sage Publications: California, 2002.
4. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
5. Doty RL, Reyes PF, Gregor T. Presence of both odor identification and detection deficits in Alzheimer's disease. *Brain Research Bulletin* 1987; **18**:597–600.
6. Devanand DP, Michaels-Marston KS, Liu X, Pelton GH, Padilla M, Marder K, Bell K, Stern Y, Mayeux R. Olfactory deficits in mild cognitive impairments predict Alzheimer's disease on follow-up. *American Journal of Psychiatry* 2000; **157**:1399–1405.
7. Doty RL, Shaman P, Dann M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiology Behavior* 1984; **32**:489–502.
8. Holland PW, Wainer H. *Differential Item Functioning*. Erlbaum: Hillsdale, NJ, 1993.
9. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth: Pacific Grove, CA, 1984.
10. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, 2000.
11. Hand DJ. *Discrimination and Classification*. Wiley: New York, 1981.
12. Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 1978; **73**:699–705.
13. Li K-C, Duan N. Regression analysis under link violation. *The Annals of Statistics* 1989; **17**:1009–1052.
14. Furnival GM, Wilson RW. Regressions by leaps and bounds. *Technometrics* 1974; **16**:499–511.

DOI: 10.1002/sim

15. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001; **29**:1165–1188.
16. Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hochberg Y. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc.: Cary, NC, 1999.
17. Tabert MH, Liu X, Doty RL, Serby M, Albers M, Zamora D, Pelton G, Marder K, Devanand DP. A 10-item smell identification scale related to risk of Alzheimer's disease. *Annals of Neurology* 2005; **58**:155–160.
18. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006; **1**:221–229.
19. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York, 2001.
20. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2004.
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristics curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
23. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
24. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; **16**:297–334.
25. Prentice RL. Use of the logistic model in retrospective studies. *Biometrics* 1976; **32**:599–606.

DOI: 10.1002/sim