

# Computerized adaptive testing with decision regression trees: an alternative to item response theory for quality of life measurement in multiple sclerosis

Pierre Michel<sup>1,2</sup>  
Karine Baumstarck<sup>1</sup>  
Anderson Loundou<sup>1</sup>  
Badih Ghattas<sup>2</sup>  
Pascal Auquier<sup>1</sup>  
Laurent Boyer<sup>1</sup>

<sup>1</sup>Aix-Marseille Univ, School of Medicine, CEReSS - Health Service Research and Quality of Life Center, Marseille, France; <sup>2</sup>Mathematics Institute of Marseille, Aix-Marseille University, Marseille, France

**Background:** The aim of this study was to propose an alternative approach to item response theory (IRT) in the development of computerized adaptive testing (CAT) in quality of life (QoL) for patients with multiple sclerosis (MS). This approach relied on decision regression trees (DRTs). A comparison with IRT was undertaken based on precision and validity properties.

**Materials and methods:** DRT- and IRT-based CATs were applied on items from a unidimensional item bank measuring QoL related to mental health in MS. The DRT-based approach consisted of CAT simulations based on a minsplit parameter that defines the minimal size of nodes in a tree. The IRT-based approach consisted of CAT simulations based on a specified level of measurement precision. The best CAT simulation showed the lowest number of items and the best levels of precision. Validity of the CAT was examined using sociodemographic, clinical and QoL data.

**Results:** CAT simulations were performed using the responses of 1,992 MS patients. The DRT-based CAT algorithm with minsplit = 10 was the most satisfactory model, superior to the best IRT-based CAT algorithm. This CAT administered an average of nine items and showed satisfactory precision indicators ( $R = 0.98$ , root mean square error [RMSE] = 0.18). The DRT-based CAT showed convergent validity as its score correlated significantly with other QoL scores and showed satisfactory discriminant validity.

**Conclusion:** We presented a new adaptive testing algorithm based on DRT, which has equivalent level of performance to IRT-based approach. The use of DRT is a natural and intuitive way to develop CAT, and this approach may be an alternative to IRT.

**Keywords:** computerized adaptive testing, binary decision trees, classification and regression trees, item response theory, quality of life, multiple sclerosis

## Background

Quality of life (QoL) measurements are increasingly being considered important for patients with chronic diseases.<sup>1,2</sup> Self-reported questionnaires are traditionally used to measure QoL, but they are often considered too lengthy by professionals whose main role is providing patient care.<sup>3</sup> In addition, questionnaires should be as brief as possible because of the difficulties of fatigue and concentration in some clinical populations, such as patients with multiple sclerosis (MS).<sup>4</sup> It is thus necessary to provide shorter questionnaires in QoL measurements.<sup>5</sup> Short-form instruments are usually a fixed length and adapted from a long-form instrument. However, these fixed-length short-form instruments have several drawbacks. The reduction in questions carries a

Correspondence: Laurent Boyer  
Aix-Marseille Univ, School of  
MEDICINE - La Timone Medical  
Campus, EA 3279: CEReSS - Health  
Service Research and Quality of Life  
Center, 27 Boulevard Jean Moulin,  
13005 Marseille, France  
Tel +33 6 8693 6276  
Email laurent.boyer@ap-hm.fr

risk of losing relevant information that can result in a decline in measurement precision.<sup>6,7</sup> Moreover, because some items are not tailored to patients, the precision of the QoL measurement is not maximized.

Methods based on item response theory (IRT) can be applied to overcome these issues. IRT parametric models are currently used for the development of unidimensional item banks and are the basis of computerized adaptive testing (CAT).<sup>8,9</sup> CAT administers only the items that offer the most information about a given individual, improving both the completion time and the measurement's precision.<sup>10-12</sup>

However, IRT is based on fundamental assumptions (ie, unidimensionality, local independence and monotonicity) that may not be met for some real datasets. In addition, IRT models require an adequate sample size depending on the number of items and response categories.<sup>13,14</sup> Moreover, the computational constraints due to several selection criteria in CAT, such as minimal expected posterior variance (EPV), may be inapplicable in practice.<sup>15</sup> An interesting alternative to IRT-based CAT may be to develop a CAT based on machine learning and decision trees (DTs); the most well-known method is classification and regression trees (CARTs).<sup>16</sup> These tree-based methods aim to get a partition of a data sample in a set of homogeneous subsamples and can be applied to both regression (decision regression trees, DRTs) and classification (decision classification trees, DCTs) problems. DCT and DRT differ according to the outcome of interest which is categorical (eg, presence of symptom [SPT] and health status) for DCT and continuous (eg, a QoL score) for DRT. Moreover, in DRT, the prediction of the response for a given observation is obtained using the mean response of the training observations in the corresponding subsample. In contrast, in DCT, the prediction of the response is obtained using the most commonly occurring class of the training observations in the corresponding subsample. The criterion to minimize to get the optimal partition of a sample is also different for the two approaches: in DRT, it is based on the residual sum of squares, while in DCT, it is based on the classification error rate.<sup>17</sup> DRT provides a simple way to represent the results obtained and does not require any model assumptions. Therefore, constructing a DRT-based CAT using an external criterion (eg, a test or dimension score) is a good way to avoid the IRT fundamental assumptions<sup>18</sup> and requires fewer computational resources.<sup>15</sup> To date, DT-based CAT has been mainly developed in the field of educational testing,<sup>19-21</sup> and several studies have reported that this nonparametric and non-iterative approach showed superior performance to IRT-based CAT, especially when

IRT assumptions were not satisfied.<sup>18,22</sup> To the best of our knowledge, the only applications of DT in medicine have been in the development of a CAT as a screening tool for depression<sup>23</sup> and a support decision system for the assessment of suicidal risk<sup>24</sup> using DCT. No study has presented a comparison between DRT-based CAT and classical IRT-based CAT in QoL measurements.

Thus, the aim of this study was to determine whether DRT-based CAT can be an alternative approach to IRT-based CAT in a single case of QoL measurements. For this purpose, we compared the performances (ie, the mean number of items administered, precision and validity properties) of DRT- and IRT-based CAT on a calibrated item bank measuring QoL related to mental health (MH) for patients with MS.<sup>25</sup>

## Materials and methods

### Study design and setting

Data from an international, multicentre, cross-sectional study were used.<sup>26</sup> All the details about this study have been taken from a previously published study.<sup>25</sup> Patients were recruited between January 2004 and February 2005 from neurology departments in 15 countries: Argentina, Canada, France, Germany, Greece, Israel, Italy, Lebanon, Norway, Russia, South Africa, Spain, Turkey, the UK and the USA. This study was performed in accordance with the Declaration of Helsinki and all applicable regulatory authority requirements and national laws. It was also conducted in compliance with the requirements of the institutional review boards and independent ethics committees of each of the 15 countries from which study participants were drawn. In France, it was a Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale ethical committee. Written informed consent was obtained from patients before any study procedures were performed.

### Population

The inclusion criteria for this study included the following: having a diagnosis of MS according to the McDonald criteria,<sup>27</sup> inpatient or outpatient, age superior to 18 years and informed consent.

### Data collection

The following data were collected:

- Sociodemographic information: gender, age, educational level, marital status and occupational status.
- Clinical data: MS subtype;<sup>28</sup> disease duration; MS disability assessed using the Expanded Disability Status Scale<sup>29</sup> (EDSS); severity of SPTs using a self-reported checklist of 14 SPTs (lack of sensation in touch, lack of sensation

in position, involuntary body movements, vibration in legs or arms, weakness in limbs, tingling in limbs, inability to swallow, involuntary eye movements, visual problems, difficulty concentrating, fatigue, urinary incontinence and bowel incontinence) for which the total score was calculated as the sum of the SPT scores (a higher score indicated more severe SPTs). This list was elaborated by an international steering committee including 15 senior neurologists, two experts in QoL, one expert in health economics and two external advisors.<sup>26</sup> The 14 items were based on the impairment in the eight functional systems of the EDSS.<sup>30</sup>

QoL was assessed using two widely used scales, the Multiple Sclerosis International Quality of Life (MusiQoL)<sup>26</sup> and the Short-Form 36 (SF-36)<sup>31</sup> questionnaires. The MusiQoL is an MS-specific questionnaire that describes nine QoL dimensions and yields a global index score. The nine QoL dimensions measured by this questionnaire are the following: activities of daily living (ADLs), psychological well-being (PWB), SPTs, relationships with friends (RFR), relationships with family (RFa), relationships with health care system (RHCS), sentimental and sexual life (SSL), coping (COP) and rejection (REJ). The SF-36 is a generic questionnaire describing eight subscales: physical function (PF), social function (SF), role physical (RP), role emotional (RE), MH, vitality (VT), bodily pain (BP) and general health (GH). Two composite scores (physical and mental: physical composite score [PCS]-SF-36 and mental composite score [MCS]-SF-36) were also calculated. Both the MusiQoL and SF-36 yield scores on a scale from 0 to 100, in which 0 represents the lowest and 100 the highest QoL scores. Both questionnaires are available in the 15 languages selected for this study.

## CAT development and simulations

DRT- and IRT-based CAT were applied on 22 items from a calibrated unidimensional item bank measuring QoL related to MH.<sup>25</sup> We implemented a real-data simulation approach, ie, complete response patterns to the 22 items were used to simulate the conditions of the CAT assessments. We used the responses contained in the item bank to simulate the adaptive administration of items.

### QoL item bank related to MH

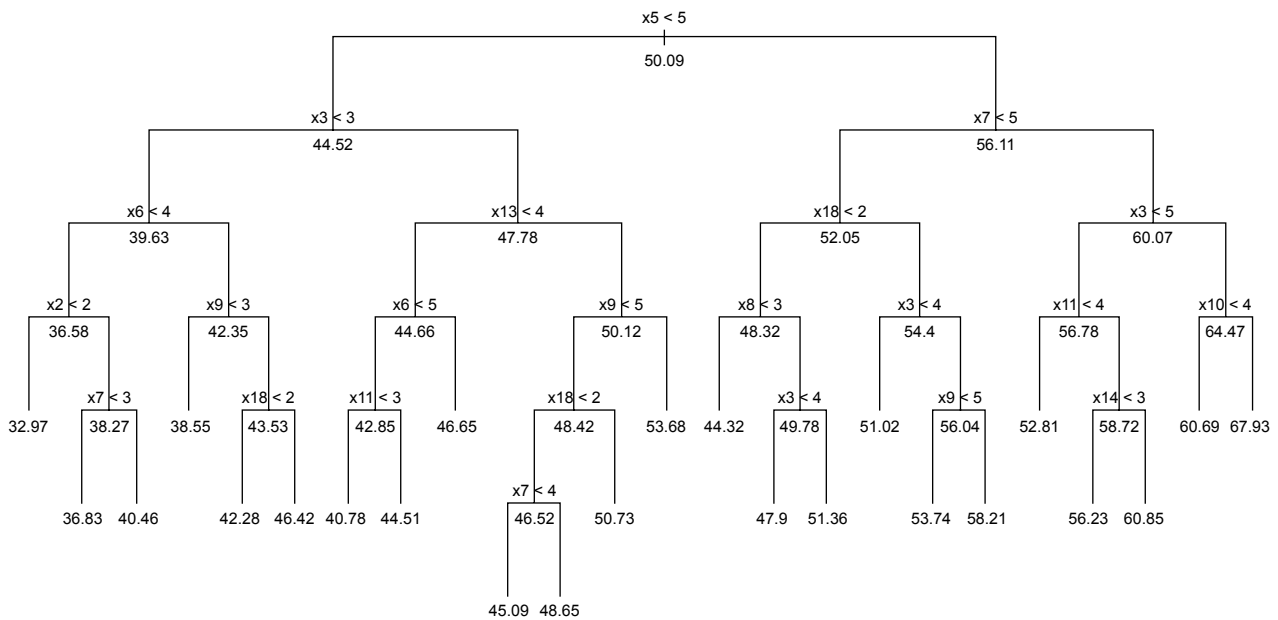
The item bank comprises 22 items issued from both MusiQoL and SF-36 questionnaires. In a recent study,<sup>25</sup> this set of items was calibrated using a partial credit model, with each item showing satisfactory “inlier-pattern-sensitive fit statistic” goodness-of-fit statistics (ie,  $0.7 < \text{inlier-pattern-sensitive fit statistic} < 1.3$ ). This study demonstrated satisfactory

psychometric properties of the item bank, and the score computed from the item bank was reliable, correlated significantly with other QoL scores and showed good discriminant validity according to sociodemographic and clinical features. The initial different dimensions represented by this set of items were PWB, COP, REJ, RE, SF and MH. Patients’ responses to these 22 items provided an item bank score, computed using maximum likelihood estimation. The score was logit-transformed to a scale ranging from 0 to 100 (with 0 representing the lowest and 100 representing the highest QoL levels). Differential item functioning (DIF) analyses showed that item biases were negligible across geographical areas, highlighting the cross-cultural equivalence of the item bank.

### DRT-based CAT

In this study, we used DRT as defined by Breiman’s CART method.<sup>32</sup>

Briefly, the items in a DRT-based model are assumed to be formative indicators, ie, items are viewed as causes of the theoretical construct under consideration.<sup>33</sup> A DRT is constructed using an iterative process, in which a binary splitting rule is searched for. For each variable  $X_j$  in the data, a splitting rule of the form  $x_j < a$  ( $a \in \mathbb{R}$  is a threshold) is used to split the initial set of observations (denoted  $t_0$ , the root of the tree) into two subsets  $t_l$  and  $t_r$  (the sibling nodes). Among the possible splits over all the input variables and all thresholds, the best split is defined as the one minimizing the sum of the within-class heterogeneity (or deviance) within the sibling nodes. Once the best split has been defined, the same process is applied to the two nodes  $t_l$  and  $t_r$  and repeated until subsets contain few observations (other stopping criteria can be considered, such as a minimum heterogeneity criterion). A predicted value is assigned to each subset of observations. Then, a pruning algorithm can be used to search for an optimal tree, given a penalized criterion (eg, deviance). A DT can be represented graphically and thus can be directly interpretable, given its simple structure. In the context of CAT, each node of the tree can be seen as a particular item of the item bank (eg, the root of the tree corresponds to the first item administered to a patient). According to the response provided by the patient, and given the threshold observed at each node, the choice of the next item is made, corresponding to the left node if the splitting rule is satisfied and to the right node otherwise. The series of items is administered sequentially, and when all the items are completed, a score is computed for the patient, represented by the predicted value in the corresponding tree node. Figure 1 shows an example of



**Figure 1** Structure of tree obtained using CART with minsplit = 100.

**Notes:** At each node, the split is written as  $x_j < a$  (top), and the mean score predicted at this node is given (bottom). Observations that satisfy the splitting rule are sent to the left child node.

**Abbreviation:** CART, classification and regression tree.

tree structure obtained from the item bank used in this study, fixing the minimal number of observations (or individuals) in a node to be split to 100 (ie, minsplit = 100).

The dependent variable  $Y$  was the item bank score. The input ordinal variables corresponded to the 22 items contained in the item bank. For tree construction, the only fixed parameter was the minsplit parameter, controlling the minimum number of observations that must remain in a tree node to split this node. The DRT-based CAT was run under four values of minsplit (100, 50, 10, 5), with lower values implying deeper trees. The CAT procedure was stopped when all items of a tree path were completed.

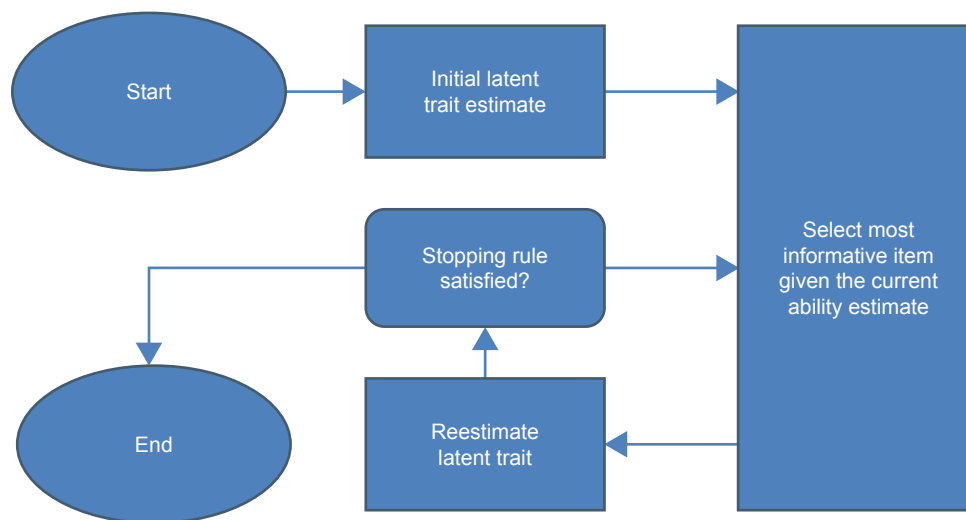
To predict the CAT score of a patient, the initial item responses can be used to follow the corresponding paths in the DT. For example, as shown in Figure 1, if a patient responded with the highest modality to items 5, 7, 3 and 10, then he/she would be sent to the extreme right node of the tree and assigned a CAT score of 67.93. Therefore, the final CAT score is defined as the predicted value of  $Y$  in the terminal node (ie, the sample mean of the dependent variable in that node), given the series of ordinal responses. We used the R package “rpart”<sup>34</sup> to apply this method.

**IRT-based CAT**

In contrast to the previous model, all the items in the IRT model were considered as reflective indicators. Following the study by Costa,<sup>35</sup> we may assume that these items were the

manifestations of an underlying construct (ie, QoL related to MH).

The item selection algorithm of the IRT-based CAT was based on the maximum Fisher information item selection (MFI),<sup>36</sup> relevant for polytomous items and adapted to an unidimensional item bank.<sup>37</sup> First, an initial latent trait estimate was provided to each patient (ie, the value was set to the study sample mean). The CAT algorithm selected the item where information function reached its maximum for the given  $\hat{\theta}$  estimate. The  $\hat{\theta}$  estimate was then updated given the patient’s response and the item parameters using Bayesian maximum a posteriori (MAP) estimation.<sup>38</sup> A new item was selected based on the new  $\hat{\theta}$  estimate and so on. The stopping rule employed was the prespecified level of measurement precision using the standard error of measurement (SEM).<sup>39</sup> A range from 0.33 to 0.55 has been selected, corresponding to reliability coefficients between 0.70 and 0.90.<sup>39</sup> For high-stakes decision making, a reliability coefficient  $>0.90$  was expected. The IRT-based CAT was run under three levels of minimally required SEM (ie, minimum, middle and maximum values: 0.33, 0.44 and 0.55, respectively), ie, the CAT procedure was stopped when the SEM threshold was reached. The last updated  $\hat{\theta}$  estimate was the final CAT score. We used the R package “mirtCAT”<sup>40</sup> to apply this method. Figure 2 shows a diagram illustrating the IRT-based CAT algorithm used in this study.



**Figure 2** IRT-based computerized adaptive test algorithm.  
**Abbreviation:** IRT, item response theory.

### Comparisons of DRT- and IRT-based CAT

For the seven simulations (ie, four for DRT-based CAT and three for IRT-based CAT), CAT scores were computed. The mean number of items was administered, and precision and validity properties were assessed for each simulation. Accuracy was assessed using the level of correlation between the CAT score and the latent trait score based on the full set of items ( $R > 0.9$  were expected). Precision was assessed using the root mean square error (RMSE). The RMSE is the square root of the mean square of all the errors. The error is the gap between the latent traits estimated by the CAT and the latent traits estimated by the full item bank. Smaller values of RMSE represent better measurement precision, and RMSE values  $\leq 0.3$  indicate excellent measurement precision.<sup>41</sup> The optimal CAT algorithm was selected based on the lowest number of items that matched with the most satisfactory levels of accuracy and precision.

A cross-validation was performed to assess the performance of our new DRT-based CAT approach and the stability of the optimal model. This analysis is presented in Table S1 and Figure S1.

### Convergent and divergent validity of the selected CAT algorithm

To explore the convergent validity of this score, relationships were investigated between the CAT score estimate and all of the scores of the MusiQoL and SF-36 dimensions. The underlying assumption was that the CAT score would be more correlated with scores of the component dimensions of the item bank than with other dimensions not captured in the item bank. The discriminant validity was determined

via comparisons of mean dimension scores across patient groups according to sociodemographic (ie, age, gender, educational level, marital status, occupational status and geographical area) and clinical (ie, EDSS score, SPT sum scale, disease duration and MS subtypes) features, using Student's *t*-tests, ANOVAs and Pearson's correlation coefficients.

## Results

### Study sample characteristics

The study sample included 1,992 patients with MS. The mean age was 42.2 (SD = 11.9) years, 578 patients (29.5%) were male, 601 (36.8%) were unemployed, 592 (35.2%) had a high educational level and 372 (21.7%) were single. The mean duration of the disease was 11.1 (SD = 8.8) years, and the median EDSS score was 3.0 (interquartile range = 3.5).

### DRT- and IRT-based CAT development and simulations

Table 1 gives the results of the seven CAT simulations.

The IRT-based CAT, based on a level of precision of  $SEM < 0.33$ , was defined as the most satisfactory CAT simulation, because this model was associated with the lowest number of items and performed the highest levels of accuracy ( $R = 0.96$ ) and precision ( $RMSE = 0.22$ ). With such a SEM threshold, the CAT score obtained is reliable to a level of 90%. In addition to these satisfactory results, this IRT-based CAT algorithm administered less than half of the initial items from the bank (on average nine items) to predict a QoL score. The other IRT-based CAT algorithms showed a lower performance. The algorithm with  $SEM < 0.44$  got an

**Table 1** Comparisons of DRT- and IRT-based approaches

Indicators	DRT-based CAT				IRT-based CAT		
	minsplit = 100	minsplit = 50	minsplit = 10	minsplit = 5	SEM < 0.33	SEM < 0.44	SEM < 0.55
Mean score	50.09	50.09	<b>50.09</b>	50.09	<b>50.00</b>	49.77	49.68
SD score	8.26	8.52	<b>8.94</b>	9.05	<b>8.64</b>	8.37	7.86
Minimum score	32.97	32.97	<b>23.4</b>	18.46	<b>15.77</b>	15.77	27.19
Maximum score	67.93	73.23	<b>78.38</b>	78.83	<b>79.56</b>	79.56	73.62
Number of items	5 (1)	6 (1)	<b>9 (2)</b>	14 (11)	<b>9 (3)</b>	5 (2)	2 (1)
Accuracy (R)	0.91	0.93	<b>0.98</b>	0.99	<b>0.96</b>	0.90	0.83
RMSE	0.33	0.28	<b>0.16</b>	0.09	<b>0.22</b>	0.35	0.45

**Notes:** Bold values represent the best models for both DRT- and IRT-based approaches. Data are presented as mean (standard deviation).

**Abbreviations:** CAT, computerized adaptive testing; DRT, decision regression tree; IRT, item response theory; RMSE, root mean square error; SEM, standard error of measurement.

acceptable precision with regard to the correlation ( $R = 0.90$ ), but the RMSE value exceeded the expected value of 0.3. Despite its low number of items administered (on average two items), the algorithm with  $SEM < 0.55$  did not show satisfactory accuracy and precision results.

Concerning the DRT-based CAT algorithms, each simulation had satisfactory accuracy with correlations higher than 0.90. On the contrary, the precision results were not satisfactory for the model with  $minsplit = 100$  ( $RMSE > 0.30$ ). In contrast, the three other DRT-based CAT algorithms showed a good level of precision. For  $minsplit = 50$ , the CAT algorithm showed satisfactory precision indicators ( $R = 0.93$ ,  $RMSE = 0.28$ ), but did not outperform the optimal IRT-based algorithm (ie,  $SEM < 0.33$ ). For  $minsplit = 10$ , the CAT administered as many items as the optimal IRT-based CAT but was more efficient in terms of accuracy and precision ( $R = 0.98$ ,  $RMSE = 0.16$ ). We decided to reject the DRT-based CAT algorithm with  $minsplit = 5$ , given its overly high number of items administered (14 items on average), which is associated with a negligible gain of measurement precision.

The DRT-based CAT with  $minsplit = 10$  was defined as the most satisfactory CAT algorithm.

### Convergent and divergent validity

The DRT-based CAT score ranged from 0 to 100, with 0 representing the lowest and 100 representing the highest level of QoL related to MH. Convergent and divergent validity results are presented in Table 2. Age was weakly negatively correlated with the CAT score. The CAT score was significantly higher for men, individuals with higher educational levels, those with jobs and those who did not live alone. No significant differences were shown according to the six different geographical areas. Regarding the clinical indices, no correlation was found with disease duration. As expected, the CAT score was negatively correlated with the

EDSS score and the SPT sum scale, and it differed according to the MS subtype, with the highest scores observed for the clinically isolated syndrome (CIS) patients and the lowest scores observed for the secondary progressive (SP) patients. According to the QoL data, the CAT score was strongly correlated with the dimension scores from both MusiQoL and SF-36 questionnaires that were used to develop the item bank.

**Table 2** Comparison of DRT-based CAT score with sociodemographic, clinical and QoL data

Characteristics	R	Mean (SD)	p-value
<b>Sociodemographic data</b>			
Age	-0.08		0.002
Gender			
Male		51.03 (8.74)	0.011
Female		49.75 (8.97)	
Educational level			
High		52.01 (8.87)	<0.001
Low		48.84 (8.80)	
Marital status			
Not alone		50.61 (9.53)	0.022
Single		49.85 (8.81)	
Occupational status			
Active		51.31 (8.97)	<0.001
Unemployed		48.04 (8.80)	
Geographical area			
South Africa		46.91 (7.92)	
Europe*		49.92 (8.88)	
Argentina		47.94 (9.11)	
North America**		50.29 (8.91)	
Eastern Middle***		51.99 (9.82)	
Russia		48.77 (6.99)	0.08
Clinical data			
EDSS score****	-0.19		<0.001
SPT sum scale	-0.43		<0.001
Disease duration	-0.03		0.313
MS subtype			
RR		50.61 (9.08)	<0.001
PP		50.31 (8.28)	
SP		48.05 (8.08)	
CIS		55.09 (10.33)	

(Continued)

**Table 2** (Continued)

Characteristics	R	Mean (SD)	p-value
<b>QoL data</b>			
<b>MusiQoL</b>			
ADL	0.55		<0.001
PWB	0.80		<0.001
RFr	0.23		<0.001
SPT	0.49		<0.001
RFa	0.26		<0.001
RHCS	0.23		<0.001
SSL	0.36		<0.001
COP	0.54		<0.001
REJ	0.51		<0.001
Index	0.78		<0.001
<b>SF-36</b>			
PF	0.36		<0.001
SF	0.65		<0.001
RP	0.52		<0.001
RE	0.59		<0.001
MH	0.87		<0.001
VT	0.79		<0.001
BP	0.45		<0.001
GH	0.57		<0.001
PCS	0.33		<0.001
MCS	0.82		<0.001

**Notes:** \*Europe includes Germany, France, Greece, Italy, Norway, Spain and UK. \*\*North America includes Canada and USA. \*\*\*Eastern Middle includes Israel, Lebanon and Turkey. \*\*\*\*The EDSS is an ordinal clinical rating scale ranging from 0 (normal neurological examination) to 10 (death due to MS).

**Abbreviations:** ADL, activity of daily living; BP, bodily pain; CAT, computerized adaptive testing; CIS, clinically isolated syndrome; COP, coping; GH, general health; DRT, decision regression tree; EDSS, Expanded Disability Status Scale; MCS, mental composite score; MH, mental health; MS, multiple sclerosis; MusiQoL, Multiple Sclerosis International Quality of Life; PCS, physical composite score; PF, physical function; PP, primary progressive; PWB, psychological well-being; QoL, quality of life; RE, role emotional; REJ, rejection; RFa, relationships with family; RFr, relationships with friends; RHCS, relationships with health care system; RP, role physical; RR, relapsing–remitting; SF, social function; SF-36, Short-Form 36; SP, secondary progressive; SPT, symptom; SSL, sentimental and sexual life; VT, vitality.

## Discussion

To the best of our knowledge, this study is the first to compare DRT- and IRT-based CAT in QoL measurements. In this study, we demonstrated that the DRT-based CAT exhibited satisfactory precision and accuracy properties, equivalent to the best IRT-based CAT. This preliminary finding confirmed the interest of considering DRT for the development of CAT in medicine, and more largely this finding opens new avenues of research and reflection on the development strategy of new QoL measures.

Our study suggests that DRT could be a relevant method for the item selection process in CAT development and an alternative to IRT, which is currently facing several unresolved problems.<sup>18</sup> In particular, IRT tends to select the items with high values of item discrimination parameters,

leading to item selection bias or errors in estimates of item parameters from an item bank. For DRT, item discrimination is defined in a nonparametric way using a deviance criterion which is computed at each node split in the tree. DRT has less item selection bias from an item bank than IRT, especially when the calibration sample is small.<sup>44</sup> The second problem of IRT is the large estimation error in the beginning of the test<sup>45</sup> conducted to select items, with optimal measurement properties at the wrong estimated ability value. New item selection algorithms (eg, mutual information measure, Kullback–Leibler information and likelihood weighted Fischer information measure) have been proposed, improving examinees' ability estimation accuracy, but these new algorithms incur a significant computational cost because they require numerical integration over ability parameters.<sup>18</sup> On the contrary, DRT proposes a simpler ability estimation approach, which consists of estimating the subsample means at each node split in the tree. In this way, the test is less computationally intensive than IRT and does not require time-consuming computations.<sup>18</sup>

Beyond these findings focusing on the item selection process, this study provided a broader reflection on the development strategy of new QoL measurements based on CAT technology. The item selection is the key step of CAT development, but the construction and the calibration of a QoL item bank are a crucial step in proposing IRT-based CAT. Indeed, IRT models require the examination of fundamental assumptions (ie, unidimensionality, local independence and monotonicity). However, the development of a QoL item bank requires substantial resources and time. It is unrealistic to develop IRT-based CAT, because it is expensive to develop, given the resources necessary to calibrate unidimensional item banks, even when the calibration time may be shortened based on a low number of items and avoidance of overly lengthy item selection stages. Pending completion of this important work (ie, development of a QoL item banks), the development of CAT based on DRT can be an attractive option because DRT does not require any assumptions or item banking procedures. DRT-based CAT can be developed from the total score of any validated scale.

We demonstrated that the optimal DRT-based CAT had satisfactory precision and accuracy properties. This DRT-based CAT algorithm administered less than half of the initial items from the bank (on average nine items). The level of correlation with the score based on the full set of items was higher than 0.9, and the RMSE was lower than 0.3. In addition, the external validity of the DRT-based CAT was consistent with our hypotheses. The DRT-based CAT scores were moderately correlated

with the EDSS according to a previous study.<sup>5</sup> The lowest DRT-based CAT scores were reported by patients with the SP form of MS, confirming that it is the most clinically aggressive and severe form of the disease. In this study, women had lower QoL scores than men, which is consistent with findings from one previous study.<sup>42</sup> Older age was significantly associated with worse scores, which is also consistent with previous findings.<sup>42</sup> Higher educational level or being in a couple was associated with higher QoL levels, as previously reported in similar cross-sectional studies.<sup>26,43</sup> As expected, the DRT-based CAT score was more correlated with the dimension scores of the SF-36, with the MusiQoL constituting the item bank.

## Limitations

A limitation in our study is that there is an endogeneity issue between the items and the full score in the DRT applications. Indeed, the items are intended to determine the QoL score, while this criterion should be an independent external variable. This is not the case in IRT methodology,<sup>46</sup> where the score does not play any role during CAT score estimation.

Despite the satisfactory results in favor of DRT-based CAT, one of the main disadvantages of DRT comes from their lack of stability. The resulting tree structure can thus be very different when the data are subject to small variations. This issue presents an important perspective for future studies. It is possible to control this stability using tree aggregation methods, the most well-known of which are bagging,<sup>47</sup> boosting<sup>48</sup> and random forests.<sup>49</sup> These methods can be used to construct higher-performing prediction models but imply a loss of interpretability. Future studies should focus on taking this disadvantage into account by testing different methods of tree aggregation to improve the results obtained in this study. Analysis of the stability of the CAT score will be undertaken based on DRT-based CAT developed from multiple resampling of the data.

Although our study accounts for a large set of potentially relevant variables for external validity, other important factors such as depression should be included in the future study.

## Conclusion

The use of DRT is a natural and intuitive way to develop CAT. Based on DRT, we have presented a new algorithm of adaptive item administration. This approach was equivalent to the more “classical” IRT-based approach when controlling some parameters of both methods. CAT simulations, as well as analysis of both measurement precision and external validity, allowed us to define an optimal algorithm to adaptively administer items measuring QoL related to MH for patients with MS.

## Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

## Author contributions

PM, LB and PA conceptualized and designed the study. PA and KB coordinated the study. PA, LB and KB performed the inclusion and clinical data collection. PM, LB and BG analyzed the data. PM, LB, KB and PA interpreted the data. PM and LB drafted and wrote the manuscript. All authors contributed toward data analysis, drafting and revising the paper and agree to be accountable for all aspects of the work. All the authors approved the final manuscript.

## Disclosure

The authors report no conflicts of interest in this work.

## References

- Mitchell AJ, Benito-León J, González J-MM, Rivera-Navarro J. Quality of life and its assessment in multiple sclerosis: integrating physical and psychological components of wellbeing. *Lancet Neurol*. 2005;4(9):556–566.
- Solari A. Role of health-related quality of life measures in the routine care of people with multiple sclerosis. *Health Qual Life Outcomes*. 2005;3(1):16.
- Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med*. 2005;60(4):833–843.
- Morris J, Perez D, McNoe B. The use of quality of life data in clinical practice. *Qual Life Res*. 1998;7(1):85–91.
- Baumstarck K, Boyer L, Boucekine M, Michel P, Pelletier J, Auquier P. Measuring the quality of life in patients with multiple sclerosis in clinical practice: a necessary challenge. *Mult Scler Int*. 2013;2013:524894.
- Walker J, Böhnke JR, Cerny T, Strasser F. Development of symptom assessments utilising item response theory and computer-adaptive testing – a practical method based on a systematic review. *Crit Rev Oncol Hematol*. 2010;73(1):47–67.
- Echteld MA, Deliens L, Onwuteaka-Philipsen B, Klein M, van der Wal G. EORTC QLQ-C15-PAL: the new standard in the assessment of health-related quality of life in advanced cancer? *Palliat Med*. 2006;20(1):1–2.
- Embretson SE, Reise SP. *Item Response Theory for Psychologists*. 1 ed. Mahwah, NJ: Psychology Press; 2000.
- Fayers P, Machin D. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. 2nd ed. Hoboken, NJ: Wiley; 2007.
- Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. *Meas Eval Couns Dev*. 2004;37(2):70.
- Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5 suppl 1):S22–S31.
- Hill CD, Edwards MC, Thissen D, et al. Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Med Care*. 2007;45(5 suppl 1):S39–S47.



13. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clin Ther*. 2014;36(5):648–662.
14. Wilson M, Engelhard G, Draney K. *Objective Measurement: Theory Into Practice*. Westport, CT: Greenwood Publishing Group; 1997.
15. Delgado-Gomez D, Baca-Garcia E, Aguado D, Courtet P, Lopez-Castroman J. Computerized adaptive tests: a decision tree approach. Reply to “Einstein, measurement, and prediction” by Gibbons et al. *J Affect Disord*. 2017;212:46–47.
16. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Boca Raton, FL: Taylor & Francis; 1984.
17. James G, Witten D, Hastie T, Tibshirani R. Tree-based methods. In: Casella G, Fienberg S, Olkin I, editors. *An Introduction to Statistical Learning*. Springer Texts in Statistics. New York: Springer; 2013:303–335.
18. Ueno M, Songmuang P. Computerized adaptive testing based on decision tree. *2010 IEEE 10th International Conference on Advanced Learning Technologies (ICALT)*; July 5–7, 2010; Sousse, Tunisia.
19. Yan D, Lewis C, Stocking M [webpage on the Internet]. *Adaptive Testing without IRT*. 1998. Available from: <http://eric.ed.gov/?id=ED422359>. Accessed November 18, 2015.
20. Yan D, Lewis C, Stocking M. *Adaptive Testing without IRT in the Presence of Multidimensionality*. Princeton, NJ: Educational Testing Service; 2002.
21. Yan D, Lewis C, Stocking M. Adaptive testing with regression trees in the presence of multidimensionality. *J Educ Behav Stat*. 2004;29(3):293–316.
22. Riley BB, Funk R, Dennis ML, Lennox RD, Finkelman M. The use of decision trees for adaptive item selection and score estimation. *Annual Conference of the International Association for Computerized Adaptive Testing*. Pacific Grove, CA, October 3–5, 2011.
23. Gibbons RD, Hooker G, Finkelman MD, et al. The CAD-MDD: a computerized adaptive diagnostic screening tool for depression. *J Clin Psychiatry*. 2013;74(7):669–674.
24. Delgado-Gomez D, Baca-Garcia E, Aguado D, Courtet P, Lopez-Castroman J. Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *J Affect Disord*. 2016;206:204–209.
25. Michel P, Auquier P, Baumstarck K, et al. Development of a cross-cultural item bank for measuring quality of life related to mental health in multiple sclerosis patients. *Qual Life Res*. 2015;24(9):2261–2271.
26. Simeoni MC, Auquier P, Fernandez O, et al. Validation of the Multiple Sclerosis International Quality of Life questionnaire. *Mult Scler*. 2008;14(2):219–230.
27. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann Neurol*. 2001;50(1):121–127.
28. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology*. 1996;46(4):907–911.
29. Kurtzke JF. On the evaluation of disability in multiple sclerosis. *Neurology*. 1961;11:686–694.
30. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. 1983;33(11):1444–1452.
31. Leplege A, Ecosse E, Pouchot J, Coste J, Perneger T. *MOS SF36 Questionnaire. Manual and Guidelines for Scores' Interpretation*. Paris: Editions Estem; 2001.
32. Breiman L. *Classification and Regression Trees*. 1st ed. Wadsworth International Group; 1984.
33. Kossakowski JJ, Epskamp S, Kieffer JM, van Borkulo CD, Rhemtulla M, Borsboom D. The application of a network approach to Health-Related Quality of Life (HRQoL): introducing a new method for assessing HRQoL in healthy adults and cancer patients. *Qual Life Res*. 2016;25(4):781–792.
34. Terry T, Atkinson B, Ripley B [webpage on the Internet]. *Rpart: Recursive Partitioning and Regression Trees*. 2015. Available from: <https://CRAN.R-project.org/package=rpart>. Accessed April 19, 2018.
35. Costa DSJ. Reflective, causal, and composite indicators of quality of life: a conceptual or an empirical distinction? *Qual Life Res*. 2015;24(9):2057–2065.
36. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. Abingdon: Routledge; 1980.
37. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. *Appl Psychol Meas*. 2009;33(6):419–440.
38. Owen RJ. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *J Am Stat Assoc*. 1975;70(350):351–356.
39. Harvill LM. Standard error of measurement. *Educ Meas Issues Pract*. 1991;10(2):33–41.
40. Chalmers RP. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *J Stat Softw*. 2016;71(1):1–38.
41. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Qual Life Res*. 2010;19(1):125–136.
42. Michel P, Baumstarck K, Ghattas B, et al. A multidimensional computerized adaptive short-form quality of life questionnaire developed and validated for multiple sclerosis. The MusiQoL-MCAT. *Medicine (Baltimore)*. 2016;95(14):e3068.
43. Fernández O, Baumstarck-Barrau K, Simeoni M-C, Auquier P; MusiQoL Study Group. Patient characteristics and determinants of quality of life in an international population with multiple sclerosis: assessment using the MusiQoL and SF-36 questionnaires. *Mult Scler*. 2011;17(10):1238–1249.
44. van der Linden WJ, Glas CAW, editors. *Elements of Adaptive Testing*. 2010 ed. New York: Springer; 2010.
45. Chang H-H, Ying Z. A global information approach to computerized adaptive testing. *Appl Psychol Meas*. 1996;20(3):213–229.
46. Edwards JR. The fallacy of formative measurement. *Organ Res Methods*. 2011;14(2):370–388.
47. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–140.
48. Freund Y, Schapire RE. Experiments with a New Boosting Algorithm; 1996. Available from: <https://people.cs.pitt.edu/~milos/courses/cs2750/Readings/boosting.pdf>. Accessed April 19, 2018.
49. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.

## Supplementary materials

### Cross-validation

To assess the performance of our new decision regression tree (DRT)-based computerized adaptive testing (CAT) approach and the stability of the optimal model, we used a resampling technique called “test set validation” which splits the data sample in two subsamples. The first subsample, containing two-thirds of the observations, was defined as the training sample, used to build the DRT-based CAT, the second subsample, containing the remaining observations, was defined as the test sample, used to validate the DRT-based CAT. This procedure was repeated for different values of minsplit ranging from 2 to 200. For each resampling, we computed the same indicators (ie, correlation and root mean square error [RMSE]) to assess the performance of our method. The

data sample was split 100 times in two subsamples for each value of minsplit: one training sample with two-thirds of the observations, one test sample with the remaining. For each test sample, we computed the same performance indicators as previously described (ie, number of items, mean score, SD and precision properties). The results were then averaged and compared to the ones obtained on the whole sample.

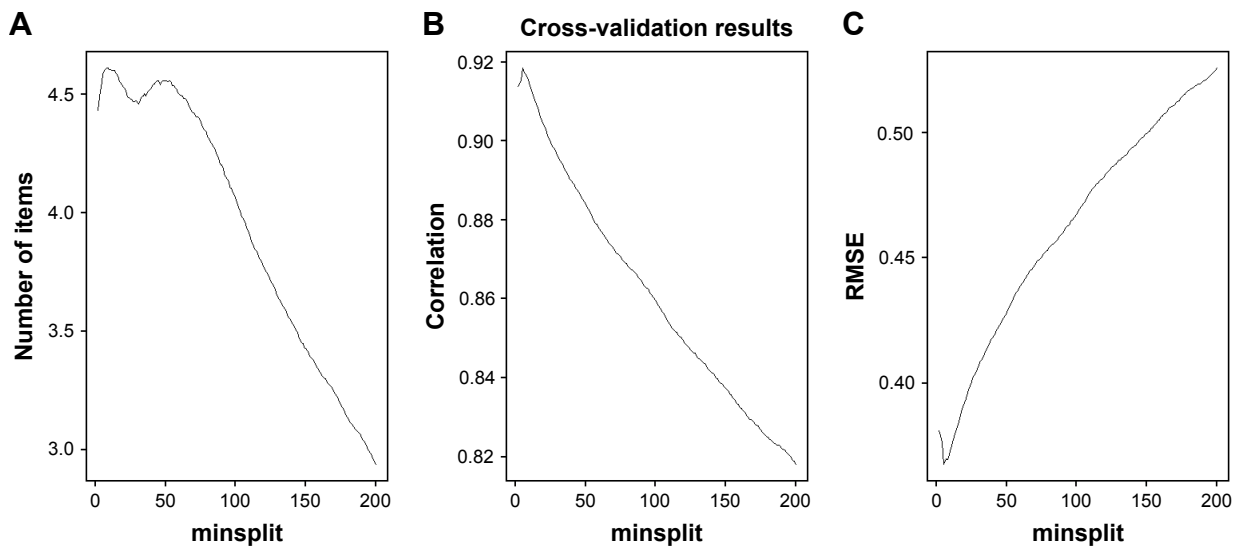
Table S1 and Figure S1 provide the results from the cross-validation. The performances of the DRT-based CAT remained stable for each value of minsplit in terms of number of items and precision. In terms of predictions, the mean score (and SD) was close to the ones computed on the full data sample. According to Table S1 and Figure S1, the most satisfactory DRT-based CAT = 10 with a lower number of items and a higher RMSE.

**Table S1** Cross-validation results for the DRT-based approach

minsplit	Mean score	SD score	Minimum score	Maximum score	Number of items	Accuracy (R)	RMSE
2	50.24	9.19	21.69	80.06	4.43	0.91	0.38
<b>10</b>	<b>50.16</b>	<b>8.84</b>	<b>26.58</b>	<b>76.88</b>	<b>4.60</b>	<b>0.91</b>	<b>0.37</b>
20	50.15	8.64	29.51	74.06	4.53	0.90	0.39
30	50.15	8.52	30.82	73.00	4.46	0.90	0.41
40	50.13	8.42	31.87	71.74	4.52	0.89	0.42
50	50.13	8.34	32.66	70.93	4.56	0.88	0.43
60	50.15	8.27	33.20	69.09	4.50	0.88	0.44
70	50.15	8.21	33.57	67.91	4.42	0.87	0.45
80	50.15	8.16	34.15	67.81	4.32	0.87	0.45
90	50.16	8.11	34.33	67.80	4.20	0.86	0.46
100	50.16	8.06	34.92	67.43	4.06	0.86	0.47
110	50.17	7.99	35.26	66.16	3.90	0.85	0.48
120	50.17	7.94	35.39	65.68	3.77	0.85	0.48
130	50.17	7.91	35.57	65.56	3.64	0.84	0.49
140	50.17	7.87	35.72	65.42	3.53	0.84	0.49
150	50.15	7.81	35.84	65.05	3.43	0.84	0.50
160	50.14	7.77	35.97	64.33	3.33	0.83	0.51
170	50.14	7.74	36.10	64.01	3.24	0.83	0.51
180	50.14	7.70	36.36	63.78	3.13	0.82	0.52
190	50.13	7.67	36.78	63.77	3.05	0.82	0.52
200	50.13	7.63	37.40	63.72	2.94	0.82	0.53

**Note:** Bold values represent the best model.

**Abbreviations:** DRT, decision regression tree; RMSE, root mean square error.



**Figure S1** Cross-validation results (1,000 replicates, 200 values of minsplit).

**Note:** (A) The number of items in function of minsplit, (B) the correlation in function of minsplit, (C) RMSE in function of minsplit.

**Abbreviation:** RMSE, root mean square error.

### Patient Preference and Adherence

## Publish your work in this journal

Patient Preference and Adherence is an international, peer-reviewed, open access journal that focuses on the growing importance of patient preference and adherence throughout the therapeutic continuum. Patient satisfaction, acceptability, quality of life, compliance, persistence and their role in developing new therapeutic modalities and compounds to optimize

Submit your manuscript here: <http://www.dovepress.com/patient-preference-and-adherence-journal>

clinical outcomes for existing disease states are major areas of interest for the journal. This journal has been accepted for indexing on PubMed Central. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Dovepress