SELECTION METHOD OF VARIABLES FROM A SET OF DYSLEXIA SCREENING TASKS IN FRENCH SCHOOL AGE CHILDREN

Guylaine Le Jan[1, 2]; Régine Le Bouquin Jeannès[1, 2]; Nathalie Costet[1, 2]; Gérard Faucon[1, 2]

[1] INSERM, U 642, Rennes, F-35000, France
[2] Université de Rennes 1, LTSI, F-35000, France
LTSI, Campus de Beaulieu, Université de Rennes 1, 35042 Rennes Cedex, France
phone: + (33) 2.23.23.62.20, fax: + (33) 2.23.23.69.17, email: guylaine.lejan@univ-rennes1.f
web: www.ltsi.univ-rennes1.fr

ABSTRACT

Dyslexia is a specific disorder of language. Researches led on dyslexia origin have conducted to multiple hypotheses and various rehabilitation treatments. In order to help in dyslexia diagnosis, a preliminary test was created. It includes the most representative dyslexia screening tasks. But, this test is too long and has too many variables. So, to define the best combination of tasks and reduce the number of variables, a method using PCA has been developed to select a first set of variables. Then, two methods (stepwise discriminant analysis and logistic regression) were applied on the pre-selected variables to identify dyslexia detection models.

1. INTRODUCTION

Developmental dyslexia affects about 5% of school age children in France. It is traditionally defined as an enduring and heavy impairment of reading ability in spite of normal intelligence and adequate educational opportunities. Dyslexics have a specific disorder of written language and can have some associated deficits like: attention deficit, visuo-attentional deficit, auditory and memory deficits. Researches led on dyslexia origin have conducted to multiple theories (*i.e*. phonological theory [1], auditory theory [2], cerebellum theory [3], magnocellular theory [4], *etc*.). These multiple theories created various diagnosis methods and treatments which are sometimes inadequate. In order to help in diagnosing dyslexia, a preliminary test containing the most representative dyslexia screening tasks has been developed in previous works [5] and tested on 56 normal readers and 28 dyslexic children (8-10 years old). The poorest reading ability in the normal children group was 18 months below the chronological age and the reading ability was on average 27 months below the chronological age in the dyslexic group. The reading level was estimated by the "Alouette test". It gives a lexical age (*i.e* reading level) in reading a test during 3 minutes. The level is evaluated by the speed and the accuracy of reading. Moreover, dyslexic children were diagnosed during a Hospital specialized consultation. They didn't present a major deficit of attention, oral language, memory, motility, visual and auditory acuity and their intellectual quotient was superior to 50 points. Nine categories of tasks are regrouped in the preliminary test: reading, writing, memory, attention, phonology, morphology, visual attention, motor and auditory tasks. For practical reasons, it must be reduced because it is too long for dyslexia screening (three sessions of 45 minutes) and for modelling reasons, the number of variables is too large. In order to reduce this test and identify the best combination of tasks to detect dyslexia, we used two methods: a stepwise discriminant analysis and a stepwise logistic regression. Before applying any method, it was necessary to eliminate some variables in preserving the most relevant variables in each category of tasks. To do this, a method of pre-selection, using a Principal Component Analysis (PCA) has been developed.

In this work, we use a PCA method to reduce the number of variables before applying a stepwise discriminant analysis and a stepwise logistic regression. The quality of the dyslexia detection models obtained is then compared.

2. VARIABLES OF PRELIMINARY TEST

All results obtained by children during the preliminary test are stored in 49 variables regrouped in 9 categories of tasks:

1) Reading tasks*:* reading of words and pseudo words: it is carried out on 4 sheets of 20 words which are grouped according to their frequency and regularity and on 2 sheets of 20 pseudo-words. The reading speed is evaluated for each sheet and the number of regularization mistakes for frequent and few frequent words is noted. So, 8 variables allow to evaluate reading tasks.

2) Memory tasks: they are composed of three tasks: forward and backward verbal span and visual span task. They give three variables: a verbal short-term memory span, work memory span and a spatial span.

3) Attention task: it is extracted from the BREV ("Batterie Rapide d'EValuation des fonctions cognitives"): children must cross out as quickly as possible all "3" placed on a test sheet during 20 and 60 seconds. Two scores give the number of "3" crossed out during respectively 20 and 60 seconds.

4) Phonological tasks*:*

- Metaphonological tasks: Four different tasks are assessed: phonemic segmentation task (segment the word in phonemes), spoonerism task (switch syllables), initial phonemic omission task (omit the first phoneme of each word presented) and task of rime judgment (find the word which does not rime with three others). These tasks return a score on respectively 16, 10, 12, 8 points.
- Phonological automatism task: it regroups speed denomination (denominate as quickly as possible a series of letters and a series of colours) and lexical discrimination (recognize if the pronunciation of two words is the same or not). So, phonological automatisms are evaluated by a

speed denomination of letters and colours, and a score of lexical discrimination on 20 points.

5) **Morphology task**: children must find a pseudo-affixed word among affixed words (example in French: recoller, *regretter*, repartir, reparler). It gives a score on 6 points.

6) Motor task: it is an extract from NEPSY ("bilan NeuroPSYchologique"): the children must execute manual motor sequences noted on 60 points and an exercise of "tapping", this one is an evaluation of digital sleight and motor speed. The speed of digital sleight is noted.

7) Visuo-attentional task: dyslexics would have difficulties in the treatment of visual information when this information is presented rapidly [6]. A partial report of letters was integrated in a program: following a central point on a computer screen, a series of 5 letters appears during 250 ms, a dash comes under one of the letters, and then the children must indicate which letter it is. For each letter position (5 positions), a score is noted on 10 points.

8) Writing task: dictation extracted from the BELEC [7]. 11 scores are retained according to the kind of mistake.

9) Auditory tasks: VOT (Voice Onset Time) tests. VOT is the time between the release of the consonant and the start of vocal fold vibration (voicing), it is measured in milliseconds. By convention, when voicing starts before the release of the consonant, VOT is negative; when voicing and consonant release happen simultaneously, VOT equals 0 ms; when voicing starts after the release of the consonant, VOT is positive. VOT quantifies the degree of phonetic voicing. The test consists in producing a continuum whose extremities are constituted of two syllables which differ by their VOT and intermediate syllables allowing linking the extremities by progressive variation of VOT. A difference of 20 ms between VOT values of two syllables is perceptible only if the syllables belong to distinct phonemic categories. For example, the syllables /ba/ and /pa/ differentiate by respectively negative and positive VOT. The production of several intermediate VOT values generates a continuum of syllables perceived like either /ba/ or /pa/. From a continuum ranging from -40 ms to 40 ms, two exercises are proposed: (*i*) identification task where the child listens to a syllable. He must indicate if he hears rather /ba/ or /pa/. This test allows to evaluate three variables: an identification slope that is calculated using a sigmoid regression, a reactivity ratio that measures the reactivity to inflexion point and a discrimination threshold that measures the perception threshold to inflexion point, (*ii*) discrimination task where two syllables are presented. The VOT difference between these two syllables is 20 ms. In this second exercise, the child must indicate if the syllables are the same or not. Normal subjects present a discrimination peak around a VOT of 0 ms. Such a peak is not recovered for children with dyslexia [8]. Moreover, predicted VOT discrimination values were calculated from VOT identification values. Then the differences, for each pair of syllables, between predicted discrimination values and observed discrimination values are noted (7 variables).

3. METHODS

3.1 Selection of variables using a Principal Component Analysis

Principal Component Analysis (PCA) is usually employed to reduce multidimensional data sets to lower dimensions for analysis [9] [10]. More rarely, PCA can be used for discriminative purpose [11]. This method allows graphic representation of information contained in a set of quantitative variables. PCA produces factors defined as linear combination of the variables. The factors are interpreted using the contribution of variables to each factor. To reduce the number of variables to be kept in our detection tool, we first used several PCA. In a first step, we pooled variables by conceptual categories of tasks and we applied a PCA within each category. Then, we selected the most contributory variables according to their relative contribution (*RCT*):

$$RCT = \frac{(factor\ loading)^2}{inertia\ of\ the\ first\ factor} \qquad (1)$$

RCT ¡Ý$1/p$, with p the number of variables in the category, is the criterion to decide that a variable is contributory. If $RCT < 1/p$, the corresponding variable is excluded for the next step of the analysis.

3.2 Discriminant analysis

A Fisher discriminant analysis including the previously selected variables by the PCA method was then implemented. A parametric analysis was chosen, under the assumptions that each group (dyslexics and normal readers) was normally distributed, the variance/covariance matrix for each group was the same and the variables were continuous. Furthermore, prior probabilities were assigned to be equal, as the costs of misclassification in both groups. Thus the discriminant function used was linear [12].

A stepwise procedure was then implemented in order to find a parsimonious model and to select the best combination of predictors. A stepwise approach is a combination between forward selection and backward estimation: starting with no predictor in the model, the first variable to enter the model is the one which maximizes a predetermined criterion. The next variable to enter is the one which maximizes the criterion, after adjustment for the previously entered variable, and so on. At each step, a variable can be removed from the model if it becomes no more significant after entering a new variable. This is the difference with a forward selection. The criterion used for the stepwise selection was the adjusted R^2.

But, some variables may be considered as non-normally distributed because they show very asymmetric distribution or they have too few modalities. Since discriminant analysis is not very robust for such distributions, we used a logistic regression.

3.3 Logistic regression

A linear logistic regression was applied on the variables preselected following the PCA. This method allowed to

include some variables which could not be considered as normal and continuous, such as the scores of work memory and morphology. These variables were considered as categorical factors and were included in logistic regression with continuous variables. So, the analysis uses a covariance model.
Stepwise AIC (Akaike's Information Criterion) approach was implemented.

3.4 Performance estimation

As the purpose was to predict group membership, the classification accuracy of the resulting functions of discriminant analysis and logistic regression was assessed through the classification matrix which compares classification groups to actual groups. The overall percentage of children correctly classified (hit ratio), the sensitivity (detection rate of dyslexic children), the specificity (detection rate of non-dyslexic children), the false-positive rate (percentage of children classified as dyslexic who were actually not dyslexic) and the false-negative rate (percentage of children classified as non-dyslexic who were actually dyslexic) were estimated using a cross validation method.

4. RESULTS

4.1 PCA within each conceptual category

For each PCA applied on the 9 categories of variables, Table 1 shows the percentage of variance explained by the first factors. They are high for all categories, greater than 40%, except for one of them (category of auditory tasks). So, the first factors have a good capacity to synthesize the concept underlying each category of variables and can be used to select the most relevant variables for modelling.

Names of variables category	Number of variables within each category	% eigenvalues of first factor
Memory	3	54.94
Attention	2	53.09
Reading	8	74.10
Phonological	7	40.45
Morphological	1	
Visuo-attentional	5	47.40
Motor	2	59.60
Writing	11	54.29
Auditory	10	20.12

Table 1 - % variance of the first factor from PCA within each category

In order to illustrate the PCA method, Figure 1 shows the coordinate's graph of the subjects according to the factors F1 and F2 obtained by PCA within the phonological tasks category.
Dyslexic subjects are labelled by a number between 501 and 531. All other points represent normal readers. The great majority of children with dyslexia is projected on the negative part of the first factor, when normal ones are projected on the positive part, except for few of them. This indicates a high descriptive and discriminant property of the first factor. All the variables of the phonological task category do not contribute equally to this factor. The selection of the most contributing variables allows to reduce the number of tasks to consider keeping the same descriptive capacity.

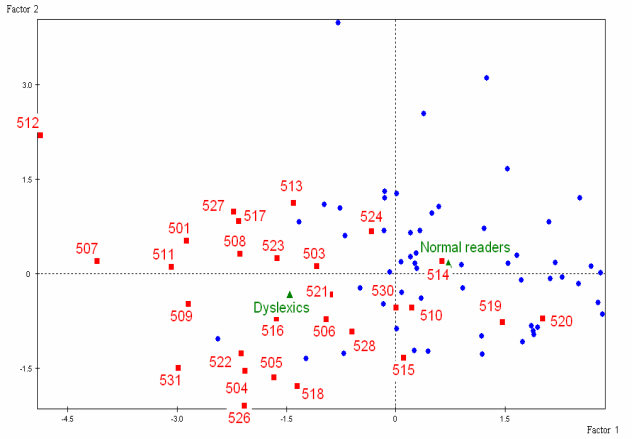

Figure 1 - Projection of individuals (children) on the first factorial plan from the PCA on phonological tasks

The analysis of the graph representing the coordinates of the variables (*cf.* Figure 2) indicates which variables are the most representative of this first factor: the best performance of all variables seem to be systematically projected on the positive part of the first factor, the worst performance being projected on the opposite part. The variables which are particularly significant on this factor are spoonerism, speed denomination and omission. Other variables (judgment of rimes, lexical discrimination and segmentation) are less convergent with the other variables to discriminate children.
Analysing conjointly children's and variable's positions on the first factor indicates that dyslexic children seem to cumulate a poor performance on different writing tasks that the PCA exhibited.

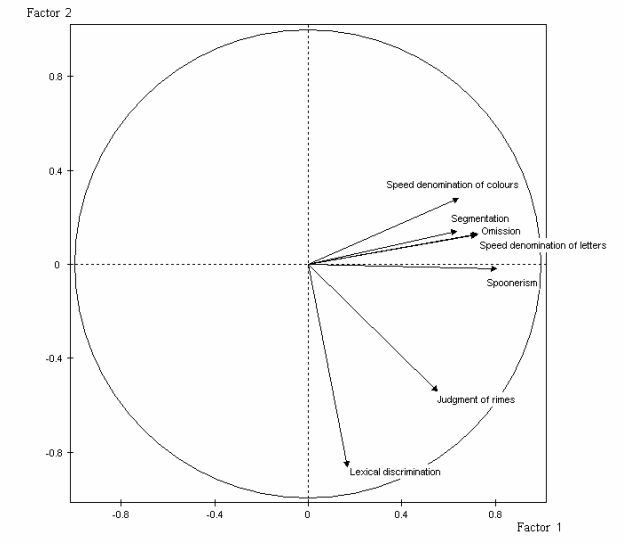

Figure 2 - Representation of phonological tasks on the first factorial plan

4.2 Contributing variables in the PCA

The selected variables using PCA method is given in Table 2. Nineteen variables ($C = 0$) have a weak influence in the construction of the first factor: visual memory span, regularization mistakes, segmentation, judgment of rimes, lexical discrimination, partial report of letters (position 3), some scores of dictation and some variables of auditory task. Using this method, we excluded variables which had a poor capacity to differentiate the two groups (dyslexic children and normal readers) and finally 30 variables were selected for the next step.

Variables names	C
Memory tasks	
Verbal forward span	1
Work memory	1
Spatial span	0
attention tasks	
3 crossing out (60s)	1
3 crossing out (20s)	1
Reading tasks	
Reading speed of frequent regular words	1
Reading speed of frequent irregular words	1
Number of regularization mistakes of frequent words	0
Reading speed of few frequent regular words	1
Reading speed of few frequent irregular words	1
Number of regularization mistakes of unfrequent words	0
Reading speed of near phonologically pseudo words	1
Reading speed of pseudo words	1
Phonological tasks	
Segmentation	0
Omission	1
Judgment of rimes	0
Spoonerism	1
Lexical discrimination	0
Speed denomination of letters	1
Speed denomination of colours	1
Morphological knowledge	
Morphology	1
Visuo-attentional task	
Partial report of letters (1 Position)	1
Partial report of letters (2 Position)	1
Partial report of letters (3 Position)	0
Partial report of letters (4 Position)	1
Partial report of letters (5 Position)	1
Motor tasks	
Manual motor sequences	1
Tapping	1
Writing task	
Dictation score 1	0
Dictation score 2	1
Dictation score 3	1
Dictation score 4	0
Dictation score 5	0
Dictation score 6	1
Dictation score 7	1
Dictation score 8	0
Dictation score 9	1
Dictation score 10	0
Dictation score 11	1
Auditory tasks	
Slope identification of speech sound	1
Ratio of reactivity	0
Discrimination threshold	1
Difference between predicted and observed values for VOT (-40 ms ; -20 ms)	0
Difference between predicted and observed values for VOT (-30ms ; -10 ms)	0
Difference between predicted and observed values for VOT (-20 ms ;+ 0 ms)	0
Difference between predicted and observed values for VOT (+0 ms ; +20 ms)	0
Difference between predicted and observed values for VOT (+10 ms ; +30 ms)	0
Difference between predicted and observed values for VOT (+20 ms ; +40 ms)	1
Difference between predicted and observed values for VOT (-40 ms ; -20 ms)	0

Table 2 - Variable selection by PCA method, $C = 1$ indicates that the variable is contributory (RCT ¡Ý $1/p$) and $C = 0$ indicates that the variable is not contributory ($RCT < 1/p$)

4.3 Comparison of dyslexia detection models by a stepwise discriminant analysis and a stepwise logistic regression

In this section, only the 30 previous selected variables were used. In a first step, a stepwise discriminant analysis was directly applied on these variables. This method gave a model including 12 variables (*cf.* Table 3).

In a second step, a stepwise AIC logistic regression was applied on the 30 selected variables. The work memory span and morphology variables have been recoded in quartile and have been considered as qualitative in the regression estimation. Six variables have been selected by this method. They are presented in Table 4.

Reading speed of frequent irregular words
Reading speed of few frequent irregular words
Reading speed of near phonologically peudo words
Spoonerism
Speed denomination of letters
Partial report of letters (position 1)
Partial report of letters (position 4)
Manual motor sequences
Dictation (Score 3)
Dictation (Score 9)
Threshold discrimination of speech sound identification task
Difference between predicted and observed values for VOT

Table 3 - Variables included in stepwise discriminant analysis model

Reading speed of near phonological pseudo-words
Spoonerism
Denomination speed of letters
Partial report of letters (position 4)
Partial report of letters (position 5)
Dictation score 9

Table 4 - Variables included in stepwise AIC logistic regression model

Figure 3 shows the predictive accuracy of each selected model. Globally, for both methods, the quality of decision rules is high (96.3% of individuals correctly classified for discriminant analysis model and 94.05% for logistic regression model). The highest sensitivity and the highest specificity are obtained using discriminant analysis model (94.82% of dyslexics and 97% of normal readers correctly classified). The logistic regression model has lower performance than the discriminant analysis but some qualitative variables could be integrated in the logistic function.

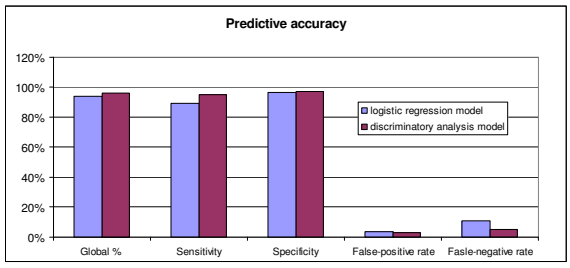

Figure 3 - Predictive accuracy (global percentage of correct classification, sensitivity, specificity false-negative rate and false-positive rate) for the two models of dyslexia detection

The clinical relevance of the selected models must be considered to choose a model among those proposed by the statistical selection procedures. The logistic regression

variables belong to four categories of tasks (reading tasks, phonological tasks, visuo-attentional tasks and writing tasks). These variables describe the most common difficulties of children with dyslexia and the tasks of this model require 20 minutes. The discriminant analysis variables belong to six categories of tasks (reading, phonological, visuo-attentionnal, writing, motor and auditory tasks). This model has the best performance and represents a lot of categories of tasks but the number of variables is too high and the tasks of this model require around 45 minutes.

5. CONCLUSION

To conclude, this study presented first a method to select variables before using a stepwise analysis. This pre-selection gives variables with the best capacity to describe both populations (dyslexics and normal-readers). Some variables as visual span memory, segmentation, judgment of rimes, lexical discrimination do not correctly characterize the dyslexics and normal-readers group. In a second step, two methods were used to select the best combination of variables able to detect dyslexics and normal-readers. The performances are generally high but the logistic regression seems more adapted to the variables and the model is relevant to clinicians. So, the preliminary test can be reduced from 2 h 45 mn to 20 mn.

REFERENCES

[1] M.J. Snowing, "Dyslexia", *Oxford : Blackwell*, 2000.

[2] P. Tallal, "Auditory temporal perception, phonics, and reading disabilities in children", *Brain Lan.*, 9, 182-198, 1980.

[3] R.I. Nicolson, A.J. Fawcett, P. Dean, "Dyslexia, development and the cerebellum", *Trends in Neuroscience*, 24(9), 515-516, 2001.

[4] J. Stein, "The magnocellular theory of developmental dyslexia", *Dyslexia*, 7, 12-36, 2001.

[5] G. Le Jan, N. Troles, R. Le Bouquin Jeannès, G. Faucon, J.E Gombert, P. Scalart, D. Pichancourt, "Développement d'une plate-forme logicielle en vue de l'élaboration d'un outil d'aide au diagnostic de la dyslexie", *Approche Neuropsychologique des Apprentissages chez l'Enfant (ANAE),* à paraître, 2008.

[6] S. Valdois, M.L. Bosse, B. Ans, S. Carbonnel, M. Zorman, D. David and J. Pellat, "Phonological and visual processing deficits can dissociate in developmental dyslexia: evidence from two case studies", *Reading and Writing: an Interdisciplinary Journal,* 16, pp. 541-572, 2003.

[7] P. Mousty, J. Leybaert, J. Alegria, A. Content and J. Morais, BELEC : une batterie d'évaluation du langage écrit et de ses troubles. Dans J. Gregoire & B. Piérat (Eds). *Evaluer les troubles de la lecture: les nouveaux modèles théoriques et leurs implications diagnostiques*, pp. 127-145, Bruxelles : De Boeck Université, 1994.

[8] W. Serniclaes and L. Sprenger-Charolles, "Categorical perception of speech sounds and dyslexia", *Special issue on language disorders and reading acquisition*, 10, vol. 1, 2003.

[9] I.T. Jollife, "Principal Component Analysis", *Springer*, 2002.

[10] F.H. Duffi, I. Valencia, G.B. Mcanulty, D.P.Waber, "Auditory evoked response data reduction by PCA : development of variables sensitive to reading disability", *ECNS, Pewaukee*, 32(3) pp. 168-178, 2001.

[11] A.M. Martinez, A.C. Kak, "PCA versus LDA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, Issue 2, pp. 228 - 233, 2001.

[12] C.J. Huberty, "Applied Discriminant Analysis", *Witey interscience*, 1994.