דוח המרכז

353

Different Approaches for Combining Scores on a Test Battery for the Diagnosis of Learning Disabilities

Tami Kennet-Cohen
Shmuel Bronner
Anat Ben-Simon
Nathan Intrator

November 2008

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

דוח מרכז 353

ISBN:965-502-144-0

Different Approaches for Combining Scores on a Test Battery for the Diagnosis of Learning Disabilities

Tami Kennet-Cohen, Shmuel Bronner & Anat Ben-Simon

National Institute for Testing and Evaluation

Nathan Intrator

School of Computer Sciences, Tel Aviv University

November 2008

Abstract

The objective of the present study was to examine alternative approaches for deriving the weights to be assigned to different performance measures in diagnosing ADD and dyslexia. These performance measures were obtained from a new standardized test battery for the diagnosis of learning disabilities (MATAL).

Two different statistical methods were applied and two definitions were used to demarcate the groups being analysed. These approaches were compared to each other on the basis of the classification accuracy of the prediction equations they yielded. The approaches were found to be highly accurate. However, this high level of accuracy should be attributed primarily to the large number and careful selection of the performance measures included in the analyses, rather than to the specific weights assigned to them by the different approaches. Additional data, especially with respect to dyslexia, is needed in order to establish any further conclusions.

Introduction

The diagnosis of learning disability (LD) is a highly complex task. In a typical clinical situation a battery of 10-20 achievement and cognitive tests is administered to a given subject and 20-40 measures may be computed to indicate performance level. To arrive at a final diagnosis, the scores on these measures – together with other measures such as medical and learning history, school reports etc. – are combined on the basis of clinical judgment. Though clinical judgment of dozens of performance outcomes may be an adequate procedure in a clinical setting (which focuses mainly on the identification of strengths and weaknesses for the purpose of designing an assistance or rehabilitation plan), it is highly inappropriate in a diagnostic setting aimed at determining eligibility for test accommodations or for financial aid, a context in which standardization and objectivity must not be compromised. Moreover, the superiority of statistical models versus clinical models in decision-making has long been established in the field (Dawes, 1979; Dawes, Faust, & Meehl, 1993).

The current study compared various approaches for constructing a statistical decision-making model for combining scores on a new standardized test-battery for the diagnosis of learning disabilities (MATAL).

MATAL is a computer-based test battery for the diagnosis of learning disabilities of students in higher education (Ben-Simon, 2005; Ben-Simon, Beyth-Marom, Inbar-Weiss, & Cohen, 2008). MATAL was developed jointly by the National Institute for Testing and Evaluation and the Council for Higher Education in Israel. One objective of MATAL is to determine eligibility for accommodations in admission tests for higher education and in university/college course exams. Another objective of MATAL is to determine the nature of the support required to assist LD students in their academic studies. MATAL consists of 20 tests in various cognitive domains. A total of 54 performance measures are derived from these tests and used to determine the existence and severity level of a learning disability. The development process of MATAL included a validation study, which was based on 205 subjects (110 subjects with one or more learning disabilities and 95 not-disabled subjects), followed by a norming study, which included 508 not-disabled subjects. The results obtained from these two studies were used to generate a statistical model (the "operational model"), which is currently used for the diagnosis of four major disabilities (dyslexia, dyscalculia, dysgraphia and Attention Deficit Disorder[1] (ADD)). The weights attached to the different performance measures in this model were estimated on the basis of a logistic ridge regression model for predicting the relevant disability. The model was optimized on the number of explanatory variables and the ridge coefficient (Wahba, 1990).

The objective of the present study was to examine alternative approaches for deriving the weights to be assigned to the different measures. These approaches differed from one another in terms of both statistical method employed and the definition given to demarcate the two groups being analysed. The analyses presented below are limited to the diagnosis of two of

[1]For the sake of accuracy it should be noted that Attention Deficit Disorder (ADD) is regarded as a cognitive disorder and not as a learning disability.

the four disabilities mentioned above, ADD and dyslexia, which are by far the most common disabilities in the current context.

Objectives of the study
(and procedures for implementing them)

Figure 1 presents a mapping sentence which describes the objectives of the study and comprises three facets. Following it is a description of each of the three facets with the elements specified in it. Methodological notes are offered where necessary to clarify technical issues regarding the implementation of the procedures designed for achieving the objectives of this study.

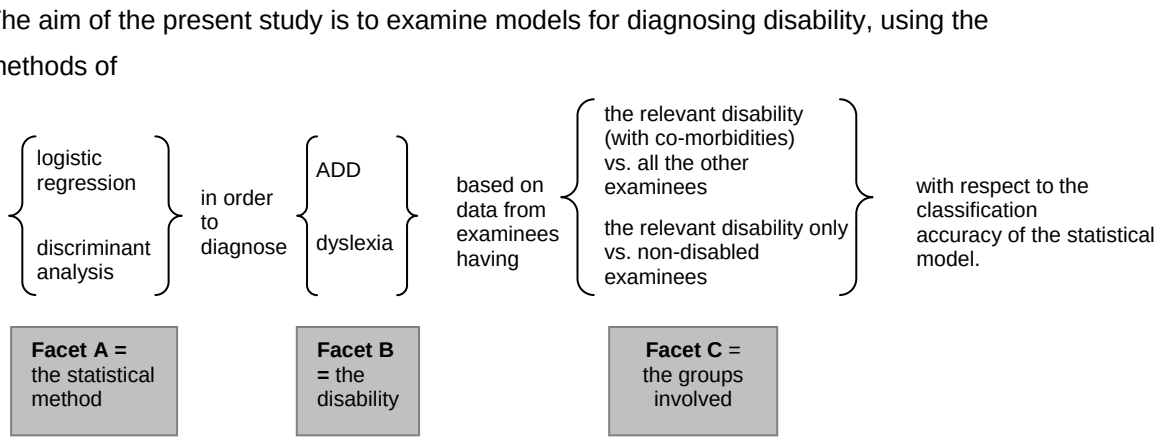

Figure 1: Mapping sentence and facet definitions for the objectives of the study

Facet A: The statistical method

In order to place examinees into categories on the basis of the battery of measurements, two statistical methods were applied: **logistic regression** and **linear discriminant analysis.** Both techniques deal with data representing multiple (continuous or categorical) independent variables (in the present application – a relevant subgroup of the performance measures of MATAL) and a single categorical dependent variable (being disabled or not, with respect to a specific learning disability[2]).

[2] The dependent variable has two categories. It should be noted that in such a two-group classification problem linear discriminant analysis is analogous to multiple regression analysis, with binary coding for the groups as the dependent variable.

Logistic regression (LR)

In **LR** we model the probability (P) that a case belongs to a particular group: $P = \frac{e^{a+b_1X_1+...+b_pX_p}}{1+e^{a+b_1X_1+...+b_pX_p}}$, with $X_1...X_p$ predictor variables. The criterion for selecting the coefficients a and $b_1,...,b_p$ is maximum likelihood, according to which, the coefficients that most closely reproduce the actual placement of cases into categories are selected.

Linear discriminant analysis (LDA)

In **LDA** we model a linear function $t = k_1x_1 + ... + k_px_p$, called a discriminant function[3]. The coefficients are selected in such a way that the scores of the members of the two categories on this function t (discriminant scores) exhibit the property of maximizing the ratio of between-groups and within-groups variability.

Thus, each of the methods yields a predicted score (a probability in the case of LR and a discriminant score in the case of LDA), on the basis of which group membership is predicted. The selection of a specific cutoff point on the predicted score for classification decisions will be discussed later.

Methodological note

The products of LR and LDA will be evaluated in reference to two alternative procedures for combining the scores on the independent variables:

1. "equal weights": combining the scores via summation with equal weights.
2. "operational weights": combining the scores via a logistic function with the operational weights (the weights which are currently used for diagnosis). The operational weights are ridge estimators applied to LR, which are obtained by maximum likelihood of a penalized model. According to Intrator (2008) "The penalty often takes the form of a ridge coefficient multiplied by the norm of the weights. Thus, the model is constrained to produce best prediction or maximum likelihood subject to a small norm constraint".

[3] t and x_i are written in lower-case letters since they are expressed as deviations of mean.

Facet B: The disability diagnosed

Each statistical analysis was conducted twice, once for diagnosing **ADD** and once for diagnosing **dyslexia**.

The dependent variable (criterion) was whether or not the respective disability was found. The independent variables (predictors) were those included in the operational equations. 17 predictors were used for diagnosing ADD and 15 for diagnosing dyslexia.

Methodological note

The decision regarding the inclusion of predictors in the operational equation was based on theoretical as well as empirical considerations. Specifically, the process began with a prediction equation which included all the 54 indices derived from MATAL. Then, indices which were irrelevant to the disability under consideration and/or appeared with small weights in the equation were omitted. An iterative process of selection of predictors followed, whereby different combinations of predictors were examined for the classification accuracy they yielded. The final aim was to obtain a maximal level of classification accuracy while retaining a theoretically acceptable combination of predictors (Ben-Simon, 2008).

Facet C: The groups involved

The models generated by the two statistical methods were estimated on the basis of two groups: disability group versus no-disability group. The elements in this facet differ in the definition used to demarcate these two groups: in the first element the two groups are defined as **"mixed"** and in the second element they are defined as **"pure."**

"Mixed" groups

The definition of the two groups as "mixed" means that the group of disabled examinees was composed of all those having the relevant disability, including cases with additional disabilities as well; and the group of not-disabled examinees was composed of "all the others" (i.e., including those with no diability at all, or those with one or more disabilities, other than the relevant disability).

"Pure" groups

The definition of the two groups as "pure" means that the group of disabled examinees was composed of cases of the relevant disability only (i.e., with no co-morbidity) and the group of not-disabled examinees was composed only of cases of no disability at all.

Methodological note

Only cases with non-missing values on the **criterion (disability**) and all the **relevant predictors** were included in the analyses.

As for the **additional disabilities** taken into account when defining the "pure" groups – dyslexia and dyscalculia were considered as additional disabilities when the criterion was ADD;

ADD and dyscalculia were considered as additional disabilities when the criterion was dyslexia.

Contrary to the treatment of cases with missing values on the predictors and the criterion, a case with a missing value on one of the additional disabilities was not omitted from the analyses. It was treated as not-disabled with respect to the disability under discussion.

The classification accuracy of the model

The models estimated by the two statistical methods in the two combinations of the groups involved were examined with respect to their **classification accuracy**.

With regard to the measure of predictive accuracy (AUC - the area under ROC curve) adopted in the present study, two issues need to be addressed. First, the rationale for this measure will be presented, introducing some key concepts from the domain of binary classification. Second, the issue of deriving this measure by means of external versus internal classification analysis will be discussed.

Key concepts: Sensitivity, specificity, ROC and AUC

The concepts of **sensitivity** and **specificity** are often used to measure the performance of a classifier.

Sensitivity is the probability that an examinee will be classified as positive (learning disabled) when he is indeed disabled; that is (true positives)/(true

positives+false negatives). The higher the sensitivity, the less real cases of disability will go undetected.

Specificity is the probability that an examinee will be classified as negative (not-disabled) when he is indeed not-disabled; that is (true negatives)/(true negatives+false positives). The higher the specificity, the less incidence of not-disabled people being labeled as disabled.

In signal detection theory, a **receiver operating characteristic (ROC)**, is a graphical plot of the sensitivity vs. (1-specificity) for a binary classifier system, as its discrimination threshold (the cutoff point) is varied.

Figure 2 shows three ROC curves representing excellent, good, and worthless predictors. The quality of the test, i.e., its discriminating power, is measured by the **a**rea **u**nder the ROC **c**urve (often called **AUC**). An area of 1 represents perfect predictor; an area of 0.5 represents worthless predictor.

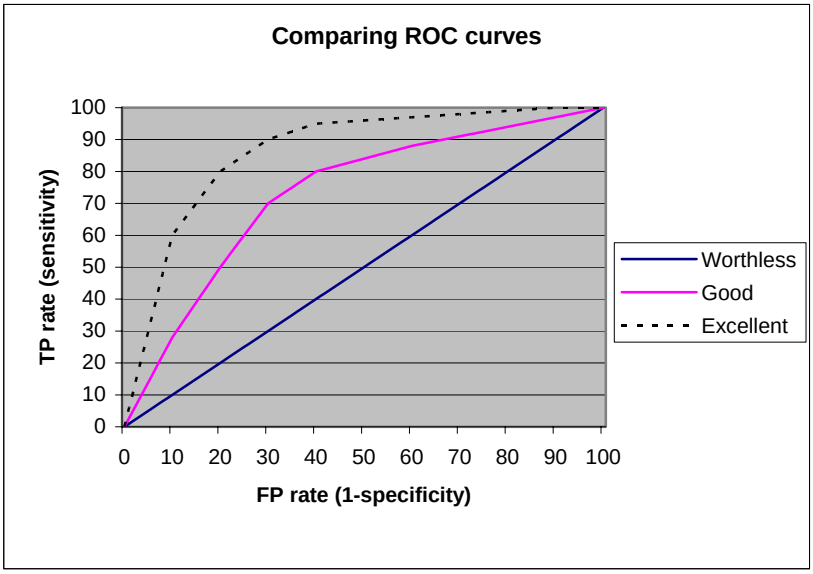

Figure 2: An illustration of ROC curves

In order to understand what **AUC** really means, consider a situation where for the validation sample (i.e., where the true condition – disabled or not – is known) we randomly pick one observation from the disability group and one from the no-disability group and examine their predicted score (i.e., the composite variable estimated by LR or LDA). The observation with the higher predicted score should be the one from the disabled group (given that the direction of the scores on the composite variable is defined such that a higher

score is associated with the disability group). **AUC** is the percentage of randomly drawn pairs for which this is true[4].

Methodological note

AUC was calculated as follows (Cortes and Mohri, 2003):

$$AUC = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{x_i > y_j}}{mn},$$

where $x_1,...,x_m$ are the predicted scores for the disabled examinees and $y_1,...,y_n$ are the predicted scores for the not-disabled examinees.

Once a satisfactory ROC curve has been obtained, the remaining task is to determine the cutoff point above which a disability is predicted and below which non-disability is predicted. There is a variety of approaches possible for determining where this cutoff point is to be located. Generally, the approach to be ultimately employed should take into account the proportion of the two groups in the population in question and the relative cost and benefits of correct and incorrect decisions.

"Internal" vs. "external" classification analysis

In internal classification analysis the same sample is used for estimating both the classification function and its accuracy. However, to obtain a realistic estimate of the predictive accuracy of a given model, external rather than internal results should be considered. External classification analysis is one in which the data to be classified are *not* used in constructing the classification function.

There are two ways of accomplishing this:

a) The *leave-one-out* (Jackknife) procedure: Here each subject is classified based on a classification function derived from the remaining (*n-1*) subjects. This is the procedure of choice for small or moderate sample sizes.

b) The *split sample* procedure: Here, the sample is randomly split (often equally) into model-building sample and validation sample. That is, we compute the classification function on the model-building sample and then

[4] It has been shown that the AUC value is equivalent to the Wilcoxon-Mann-Whitney statistic (Hanley & McNeil, 1982).

check its hit rate on the validation sample. This procedure is suitable when the sample size is large.

The samples available for the present study seemed rather small (especially the one representing the disabled group), therefore the *leave-one-out* procedure seemed more suitable. However, we decided to treat this issue with extra care, since applying the *leave-one-out* procedure to a not-too-small sample can result in a misleading optimistic picture regarding the external validity of the classification function. Therefore, the *split sample* procedure was applied in addition to the *leave-one-out* procedure.

Results

The numbers of observations on which the results are based are presented in Table 1.

Table 1

Number of observations in the analyses for ADD and dyslexia in the two combinations of groups involved

	Mixed groups		Pure groups	
	Disabled	Not-disabled	Disabled	Not-disabled
ADD	40	609	28	590
dyslexia	24	624	13	591

The values of AUC obtained by LR and LDA models are presented in Tables 2a (for ADD) and 2b (for dyslexia). Three estimates of AUC are presented for each statistical method: one obtained by internal classification analysis and two obtained by external classification analysis. The AUC values obtained by the two alternative procedures for combining the scores on the independent variables - "equal weights" and "operational weights" - are presented as well. All the AUC values mentioned above are presented for both the "pure" and the "mixed" groups.

Table 2a

Predicting ADD:

AUC values obtained through LR & LDA and through operational & equal weights in "mixed" and "pure" groups

		Mixed groups		Pure groups	
		LR	LDA	LR	LDA
Internal analysis (optimal weights)		0.971	0.969	0.982	0.979
External analysis (cross validation)	Leave-one-out	0.911	0.946	0.946	0.962
	Split sample[1]	0.920	0.942	0.913	0.947
Operational weights		0.958		0.978	
Equal weights		0.927		0.913	

[1]The values presented here are average values across the two halves of the sample. The results for each half of the sample separately are presented in the Appendix.

Table 2b

Predicting dyslexia:

AUC values obtained through LR & LDA and through operational & equal weights in "mixed" and "pure" groups

		Mixed groups		Pure groups	
		LR	LDA	LR	LDA
Internal analysis (optimal weights)		0.995	0.986	0.996	0.991
External analysis (cross validation)	Leave-one-out	0.987	0.975	0.754[2]	0.978
	Split sample[1]	0.981	0.960	0.931	0.957
Operational weights		0.981		0.983	
Equal weights		0.965		0.959	

[1]The values presented here are average values across the two halves of the sample. The results for each half of the sample separately are presented in the Appendix.
[2]The validity of model fit is questionable. There are several data points which interfere with an optimal application of maximum likelihood estimation.

Summary and conclusions

The aim of the study was to examine the quality of the diagnosis of two learning disabilities – ADD and dyslexia – using two alternative statistical methods – LR or LDA – and two alternative compositions of the groups – "mixed" or "pure" – to which the statistical analyses were applied. The quality of the diagnosis was examined on the basis of its classification accuracy via external analysis. In what follows, the main findings with respect to the above

dimensions will be reviewed, with additional comments regarding the two alternative procedures – "equal weights" and "operational weights" – for combining the scores on the independent variables.

The quality of the study[5] weights

Predicting ADD

With respect to the **statistical method**, LDA yielded somewhat higher classification accuracy. This slight superiority of LDA should be evaluated in a context of inconsistencies in the literature which compared LR and LDA (Meshbane & Morris, 1996). LDA assumes, in most applications, multivariate normality of the predictors and equal population co-variances among the predictors for the groups (Huberty, 1994). LR, on the other hand, does not make these assumptions (Hosmer & Lemeshow, 2000). When the model's assumptions are satisfied LDA is often recommended. When assumptions underlying the LDA procedure do not hold, LR is often recommended. However, as mentioned above, the issue of the relative standing of LDA and LR regarding classification accuracy is still a matter of study and discussion (Lei & Koehly, 2003).

With respect to the **groups involved** no clear-cut advantage to either approach was revealed. That being said, a slight advantage, in terms of classification accuracy, resulting from using "pure" groups, can be pointed out. In addition to this albeit limited advantage of the "pure" groups approach, there is another perspective that might be augmented. It can be argued that the approach of utilizing "pure" groups is theoretically more correct, and thus empirically more promising in practical applications. This claim stems from the fact that obtaining the weights from a sample that includes cases with co-morbidity (i.e., both ADD and dyslexia) can cause an artificial correlation between the predicted score for ADD and the predicted score for dyslexia. In other words, when a predicted score is computed based on a sample with "mixed" groups, the resulting correlation between the predicted score for ADD and the predicted score for dyslexia is higher than the correlation which exists between the two disabilities in the population. To substantiate this point we

[5] Defined by the dimensions presented in the mapping sentence in Figure 1.

compared the correlation between the predicted score for ADD and the predicted score for dyslexia in the two cases: when the weights were estimated in "mixed" groups and when they were estimated in "pure" groups. Both correlations were computed in the sample of (580) cases of no disability at all. The correlations obtained were 0.33 (when the weights were estimated in "mixed" groups) compared with 0.23 (when the weights were estimated in "pure" groups) when LDA was applied. The parallel values when LR was applied were 0.09 and 0.04 for "mixed" and "pure" groups respectively. Thus there was support[6], albeit weak, for the claim that computing the weights for the predicted score in a sample with co-morbidity yields an artificial correlation between the predicted propensity to have ADD and the predicted propensity to have dyslexia. This claim underscores the superiority of the "pure" groups approach in terms of classification accuracy.

Predicting dyslexia

Contrary to the picture obtained when predicting ADD, no substantial advantage, in terms of classification accuracy, was gained by using differential weights for the predictors when predicting dyslexia.

This conclusion is derived from the fact that half of the AUC values obtained when the weights were computed according to the dimensions examined in this study – the statistical method and the composition of the groups involved – were lower than those obtained when using equal weights. This is not surprising given the extremely small number of observations (especially when using "pure" groups) involved in predicting dyslexia. In light of these results, any attempt to compare between the statistical methods or the combinations of the groups involved might lead to erroneous conclusions.

The quality of equal weights

The most salient finding to emerge from this study is the fact that using equal weights results in an extremely high accuracy level, leaving almost no room for improvements via differential weights. This finding can be attributed to the

[6] The validity of this conclusion is based on the assumption that a given disability is the same whether an additional morbidity exists or not.

large number of **carefully selected** predictors used. Two points should be raised with respect to this finding. First, even if the large number of predictors in the prediction equations is to be kept for future use with MATAL, it is clear that the quality of the selected subset of predictors, as well as of the different schemes of weights applied to them, needs to be examined in a sample different from the one in which the predictors were selected. Second, it might be useful to examine the possibility of reducing the number of predictors, thereby gaining a more efficient process.

The quality of the operational weights

The AUC values obtained by the application of the operational weights were lower than the AUC obtained through internal analysis, as expected. It should be noted, however, that the fact that they were higher than the AUC obtained through external analysis cannot be seen as evidence of their quality, since they were not subjected to such an analysis in the current study.

References

Ben-Simon, A. (2005). *A computer-based battery for the assessment of learning disabilities in higher education*. Paper presented at the meeting of the Association of Test Publishers, Scottsdale, AZ.

Ben-Simon, A., Beyth-Marom, R., Inbar-Weiss, N. & Cohen, Y. (2008). *Regulating the diagnosis of learning disability and the provision of test accommodations in institutions of higher education.* Paper presented at the annual meeting of the International Association for Educational Assessment, Cambridge, UK.

Ben-Simon, A. (2008). *Personal communication*.

Cortes, C., & Mohri, M. (2003). AUC optimization versus error rate minimization. In *Advances in neural information processing systems, Vancouver, Canada*: The MIT Press.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.) *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351–367). Hillsdale, NJ: Erlbaum.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143(1)*, 29-36.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Intrator, N. (2008). *Personal communication*.

Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *The Journal of Experimental Education, 72(1)*, 25-49.

Meshbane, A., & Morris, J. D. (1996). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at the annual meeting of the American Research Association, New York, NY.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

Appendix

Results of the split sample procedure

Table 3a

Predicting ADD:

AUC values obtained through LR and LDA in the two halves of the sample on the basis of "mixed" and "pure" groups

	Mixed groups				Pure groups			
	LR		LDA		LR		LDA	
	1st sample	2nd sample	1st sample	2nd sample	1st sample	2nd sample	1st sample	2nd sample
For optimal weights	0.983	0.981	0.973	0.971	0.994	0.994	0.982	0.987
For split sample procedure	0.935	0.904	0.954	0.929	0.947	0.878	0.959	0.935

Table 3b

Predicting dyslexia:

AUC values obtained through LR and LDA in the two halves of the sample on the basis of "mixed" and "pure" groups

	Mixed groups				Pure groups			
	LR		LDA		LR		LDA	
	1st sample	2nd sample	1st sample	2nd sample	1st sample	2nd sample	1st sample	2nd sample
For optimal weights	1.00	1.00	0.988	0.992	1.00	1.00	0.991	0.993
For split sample procedure	0.981	0.980	0.970	0.950	0.968	0.893	0.934	0.980