NEW APPROACHES FOR IMPROVED QUALITY IN EDUCATIONAL ASSESSMENTS: USING AUTOMATED PREDICTIVE SYSTEMS IN READING AND MATHEMATICS

Mariel Musso
Leiden University, Netherlands; Assessment Group International, USA
E-mail: agi_group@msn.com

Eduardo Cascallar
Leiden University, Netherlands; Catholic University of Leuven, Belgium;
Assessment Group International, USA
Email: cascallar@msn.com

Abstract

Education has been impacted by the shift from an industrial society to an information-based environment. We are now shifting again to an "innovation-based" society which requires what Sternberg (2000) calls "successful intelligence". As the practice of educational assessment evolves, developments in cognitive science and psychometrics along with continuing advances in technology lead to new views of the nature and function of assessment (Dochy, Segers & Cascallar, 2003; Braun, 2005). Mathematics and reading have been highlighted as crucial indicators of quality in education providing essential knowledge tools and constituting the foundations for lifelong learning skills (European Report on the Quality of School Education, 2000). New methodologies and technologies, and the emergence of predictive systems, have focused on the possibility of assessments which use a wide range of data or student productions to evaluate their performance without the need of traditional testing (Boekaerts & Cascallar, 2006). This article presents the application of educational assessments utilizing neural network predictive systems in two pionneering studies in reading readiness and mathematics performance. It introduces the application of these methodologies in education, and evaluates the results and quality of the predictive systems. Results from these methods achieved excellent levels of predictive classification. Their impact on educational quality and improvement, as well as accountability is highlighted.

Key words: *assessment, mathematics education, neural networks, predictive systems, reading readiness.*

Introduction

Education has been impacted by the shift from an industrial society to an information-based environment. We are now shifting again to an "innovation-based" society which requires what Sternberg (2000) calls "successful intelligence". As the practice of educational assessment evolves, developments in cognitive science and psychometrics along with continuing advances in technology lead to new views of the nature and function of assessment (Dochy, Segers & Cascallar, 2003; Braun, 2005). In recent years, technology and the emergence of new models have had an enormous impact in improving both the quality and the utility of assessment. New technology-driven infrastructures have contributed to the quality of assessment systems, and the significant increase in available computing power together with the introduction of affordable high-speed data networks has changed many long-held assumptions for the design and the delivery of tests.

New applications have continuously been introduced which affect all aspects of the assessment process: knowledge base management, development of test items, computer delivery, and automated scoring. Currently, these advances cover a wide range of new applications using diverse technological advances, which result in the implementation of programs with novel technical and conceptual contributions. These new methodologies and technologies, and the emergence of predictive systems, have focused on the possibility of assessments which use a wide range of data or student productions to evaluate their performance without the need of traditional testing (Cascallar, Boekaerts & Costigan, 2006; Boekaerts & Cascallar, 2006)

These new tools should be sensitive enough to accrue information about the level of performance that the students have reached so far in the domain of study. This approach should also include the prediction of the expected outcomes that best capture the students' current level of learning, using already available information. There are a series of predictive approaches that examine in detail the multiple elements involved in these phenomena. In particular these series of studies use a predictive systems approach, utilizing artificial neural networks (NNs) in a stream analysis to accomplish this goal. Conceptually, a neural network is a computational structure consisting of several highly interconnected computational elements, known as neurons, perceptrons, or nodes. Each neuron carries out a very simple operation on its inputs and transfers the output to a subsequent node or nodes in the network topology (Specht, 1991). Neural networks exhibit polymorphism in structure and parallelism in computation (Mavrovouniotis & Chang, 1992), and it can be construed as a highly connected structure of processing elements that attempts to mimic the parallel computation ability of the biological brain (Grossberg, 1980, 1982; Rumelhart, Hinton, & Williams, 1986; Rumelhart, McClelland & the PDP research group, 1986).

Predictive streams analyses (Cascallar & Musso, 2008), based in this case on neural network (NN) models, have several strengths: (a) because these are machine learning algorithms, the assumptions required for traditional statistical predictive models (e. g., ordinary least squares regression) are not necessary. As such, this technique is able to model nonlinear and complex relationships among variables. NNs aim to maximize classification accuracy and work through the data in an interactive process until maximum accuracy is achieved, automatically modeling all interactions among variables; (b) NNs are robust, general function estimators. They usually perform prediction tasks at least as well as other techniques and sometimes perform significantly better (Marquez, Hill, Worthley & Remus 1991); (c) NNs can handle data of all levels of measurement, continuous or categorical, as inputs and outputs. Because of the speed of microprocessors in even basic computers, NNs are more accessible today than they were when originally developed.

The NN learns by examining individual training case, then generating a prediction for each testing case, and making adjustments to the weights whenever it makes an incorrect prediction. Information is passed back through the network in iterations, gradually changing the weights. As training progresses, the network becomes increasingly accurate in replicating the known outcomes. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met. A minimum level of accuracy can be set as the stopping criterion, although additional stopping criteria may be used as well (e. g., number of iteration, amount of time). Once trained, the network can be applied to future cases (validation or holdout sample) for validation and implementation (Lippman, 1987).

Neural Networks in Educational Research

NNs have been used in several different fields of research and in applied environments, such as: biology, business, finance, medicine, meteorology, environmental studies, and in the prediction of terrorist attacks, among other applications. During the last few decades, NNs have been increasingly utilized as a statistical methodology in applied areas such as classification and recognition of patterns in business and the social sciences (Al-Deek, 2001; Neal & Wurst, 2001; Nguyen & Cripps, 2001; White & Racine, 2001; Laguna & Marti, 2002; Detienne, Detienne & Joshi, 2003).

However, the literature shows very few studies applying neural networks in education and in educational assessment in particular (Everson, Chance, & Lykins, 1994; Wilson & Hardgrave, 1995), even though some authors have called attention to the fact that traditional statistical methods do not always yield accurate predictions and/or classifications (Everson, 1995). Preliminary research applying artificial intelligence computing methods to problems of prediction, selection and classification (Weiss & Kulikowski, 1991; Perkins, Gupta, Tammana, 1995) suggests that artificial neural networks and other neural computing methods may substantially improve the validity of the classifications, as well as increase the accuracy of classifications, and also improve the predictive validity of test scores and other educational information (Everson, Chance, & Lykins, 1994). Another study (Hardgrave, Wilson, & Walstrom, 1994) compared a neural network model to other techniques in predicting graduate student success. They evaluated five different models: least squares regression, stepwise regression, discriminant analysis, logistic regression, and neural networks. Results of their study showed that neural networks "perform at least as well as traditional methods and are worthy of further investigation" (p. 249). Similarly, Gorr (1994) used neural networks to model the decision-making process of college admissions. Neural networks were compared with linear regression, stepwise polynomial regression, and an index used by the graduate admissions committee. These researchers found that "…a neural network identifies additional model structures over the regression models" (p. 17), and that even though a neural network model can address some of the same research issues as a conventional regression, a neural network is inherently a different mathematical approach (Detienne, Detienne & Joshi, 2003). In terms of their application, neural networks have been considered to be especially good as statistical models when the emphasis is on prediction and/or classification of complex phenomena rather than on explanation (Duliba K.,1991; Bansal, Kauffman & Weitz, 1993); furthermore, recent developments have provided tools to look into the "black box" of the network, with procedures such as C5.0, and begin to shed light on the interrelationships of the variables involved in the network calculations.

In the context of these new models currently being considered for new educational assessments, the areas of mathematics and reading have been highlighted as crucial indicators of quality in education providing essential knowledge tools and constituting the foundations for lifelong learning skills (European Report on the Quality of School Education, 2000). For this reason, although the approach suggested could be generalized to other fields, the current studies center on the assessment of these two basic skills.

Measures to Evaluate the Neural Network System Performance

In order to evaluate the performance of the neural network system, there are a number of measures used which provide a means of determining the quality of the solutions offered by the various network models tried. The traditional measures include the determination of actual numbers and rates for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) outcomes, as products of the NN analysis. In addition, certain summative evaluative algorithms have been developed in this field of work, to assess overall quality of the predictive system.

These overall measures are: Recall, which represents the proportion of correctly identified targets, out of all targets presented in the set, and is represented as: Recall = P/(TP + FN); and Precision which represents the proportion of correctly identified targets, out of all identified targets by the system, and is represented as: Precision = TP/(TP + FP). Two other measures have been used to report the characteristics of the detection sensitivity of the system. One of them is Sensitivity (similar to Recall: the proportion of correctly identified targets, out of all targets presented in the set), and which is expressed as Sensitivity = TP/(TP + FN). The other is Specificity, defined as the proportion of correctly rejected targets from all the targets that should have been rejected by the system, and which is expressed as Specificity = TN/ (TN + FP). These measures are typically represented in what is called a "confusion matrix" representing all four outcomes (see Table 1).

Table 1. Confusion matrix to evaluate the outcomes of a predictive system comparing the results predicted by the system with external independent criteria or objective evidence.

		PREDICTIVE CLASSIFICATION *(by predictive system)*	
		TRUE (T)	**FALSE (F)**
OBJECTIVE CLASSIFICATION *(by independent criteria)*	**T**	**TP**	**FN**
	F	**FP**	**TN**

In addition, the evaluation of NN performance is carried out with a summative measure, which is used to account for the somewhat complementary relationship between Precision and Recall. This measure is defined as F1, and is defined as F1 = (2 * Precision * Recall)/ (Precision + Recall). Such a definitional expression of F1 assumes equal weights for Precision and Recall. This assumption can be modified to favor either Precision or Recall, according to the utility and cost/benefit ratio of outcomes favoring either Precision or Recall for any given predictive circumstance.

In order to maximize resources while maintaining or even improving the effectiveness of the neural network model in the prediction of outcomes, several methods have been explored to reduce the number of input variables needed to achieve similar predictive efficacy. One such method is the use of Item Response Theory (IRT) methods to study the functioning of each predictor variable, and its contribution to the information load for the construct being predicted (Cascallar, 2003). IRT is a set of probabilistic models which can describe the relationship between a respondent's (e. g., child) magnitude on a construct (a. k. a. latent trait; e. g., extraversion, cognitive ability, affective commitment), to his/her (its) probability of a particular response to an individual variable. This relationship can be interpreted as the informational content provided by the variable to the prediction of the construct. If a three-parameter IRT model is used, the discrimination, probability of response, and informational content can be determined, while controlling for the pseudo-random probability of a response not due to the construct assumed for the variable. This method of data reduction for input into the NN model has been shown to even improve the predictive performance of the network, by eliminating certain variables based on a variety of theory-based models. Sometimes there is a concern about the number of cases available for training purposes, but case generation methods and alternative means of training (i.e., one-out methods) are available, and applicable to problems and situations in which large data sets area not readily obtainable.

Two Implementations of Predictive Systems the Assessment of Reading and Mathematics

As it has been described, neural networks are now a well established analytical methodology. In this instance, two specific areas are studied: predictive classification of levels of reading readiness, and level of performance in a multiple-choice mathematics test. These are good examples that examine not only two completely diverse academic fields, but also two very different age groups among the students participating. The purpose is to show the applicability and effectiveness of the approach to a variety of conditions in educational assessment, as well as to understand the specific issues related to the two studies.

Study 1: Predictive Classification in Reading Readiness Research

The purpose of the study was to develop a predictive classification of students with sufficient precision to enable their accurate grouping into three groups corresponding to three levels of readiness for reading instruction. The specific task was to model the group assignment of students, using a neural network approach. This methodology, using a predictive system, was chosen as it is extremely effective under conditions of a very complex and large universe of data, in which a large number of variables interacts in various complex and not well understood patterns.

Objective: The purpose of the study was the school district's desire to avoid costly and/or time consuming individual assessments of each student, and to have the means to develop and "early warning" system which would allow early and prompt intervention with those students most in need of support and remediation. It was assumed that early intervention, before the negative experience of failure in school work would improve the possibility of achieving more successful outcomes for those identified "at risk" students, and would be an important tool for better planning and management of school resources.

Methodology of Research (Study 1)

Subjects

A total of 1052 students participated in this study (48% females). All participating students were beginning first grade in public primary schools in a large school district in the USA. The average age was 6 years. Each one of the participating students had already been classified into one of the reading readiness groups by the corresponding teacher, following the usual procedures in the district. These levels were: Level 1 ("Needs support"), Level 2 ("Adequate"), and Level 3 ("Superior"). The proportion of students in each of the levels was as follows: Level 1: 46.1%; Level 2: 33.6%; Level 2: 20.3%.

Model

The procedure used was a back propagation network, that is, a multilayer network composed of nonlinear units, which computes its activation level by summing all the weighted activations it receives and which then transforms its activation into a response via a nonlinear transfer function. Following this process, the network learns nonlinear mappings between pairs of input-output patterns to produce a predictive model for the dependent variable (group classification), based on the values of the predictor variables. Such a network is used as a pattern classifier to solve nonlinear problems. It uses supervised learning, so that the difference between the response of an output unit (the output classification) and the expected response is the error made by the network, which is the basis for connection weight correction. The units in the output layer use the error directly to correct their connection weights. The units of the hidden layer, since they are not in direct contact with the error, need to estimate it. In order to achieve this, the amount of error is converted into an error signal that is in proportion to the rate of change (the derivative of the nonlinear transfer function). Once obtained, the error signal is transferred backwards ("error backpropagation") through the connection weights to the hidden layer. The units in the hidden layer then estimate their error as a weighted sum of the error signals received from the connected output layer units. Once the error signal from the units in the hidden layer has been estimated, then the connection weights throughout the network are updated proportionally to their error signal. This method for the training of the network is one of the most widely used in various fields, as it has been shown (Haykin, 1998) that backpropagation networks are "universal approximators". That is, any arbitrary mapping between input and output can be modeled and approximated, if the correct architecture is used. The process determines, for a given input vector (predictors), the probability that it belongs to a certain class or category, given its value and learning history. In other words, in our studies, a given element of the response vector is an approximation of the *a posteriori* class probability.

Architecture of the neural network

Approximately 480 variables were initially included in the study. These variables include information in a wide range of areas, such as basic background information of each student, family system, socio-economic data, level of education of parents, occupation of family members, schooling information of the students, characteristics of the community, reading habits of family members, computer and internet use in the family, parents' attitudes, and other similar variables. Several questionnaires, as well as information provided by the schools were used for the development of the vector-matrix containing all predictor variables for each student. The resulting network contained all the input predictors, some of them collapsed into subscales to maximize predictive classification, one hidden layer, with 13 units, and three output categories. A standardized method for the rescaling of covariates was used. The hidden layer activation function was a hyperbolic tangent. For the output layer, the activation function chosen was identity, and the error function the sum of squares.

Procedure for network development

The software used was Clementine (v. 11) for the development and analysis of all predictive models in this study. One of the first and most important tasks was the consolidation of the variables into subscales, either for theoretical or data-driven reasons. Each subscale had to have an internal consistency minimum (alpha equal or greater than .75). The sample of students was divided into three groups in order to carry out the three development phases of the predictive system: training of the network, testing of the network developed, and validation of the network. During the training phase several models were attempted, and several modifications of the neural network parameters were tried, manipulating learning persistence, learning rate, momentum, and other criteria. These tests continued until achieving desired levels of classification, maximizing the benefits of the models chosen. In this analysis both precision and recall, as outcome measures of the network, were given equal weight. During the training process, several methods of trimming of predictors were used, including discrimination analysis, decision trees, and methods derived from item-response-theory (IRT). These methods were applied to the final vector matrix in order to obtain an optimum number of final complex predictor variables which maximized the predictive classification of the system. Thus, three final models were tested: Model 0, with 48 complex predictors; Model 1, with 16 complex predictors, and Model 2, with 25 complex predictors. During the testing phase, the network receives the outcome information (the actual classification of each participating student in the corresponding level of performance. During the testing phase, the network does not receive this information and performs the classification based on the models developed in the training phase, on a new vector matrix containing the predictor information for a new group of students. During both phases of the study the final confusion matrix for the neural network was determined, to evaluate its performance for each model developed.

Results of Research (Study 1)

The results obtained were very successful in terms of the effectiveness of the predictive classificatory models developed. After various iterations and testing of alternative architectures three final models were developed (see Figures 1–3). Model 1 was the most successful and obtained the best predictive classification for students at all levels of reading readiness. The models were developed to maximize the accuracy for the detection of students in the lowest expected performance level ("needs support"), coinciding with the objectives of the study. The true positives (those students correctly identified as belonging to the category in question: "needs support", "adequate" or "superior" in terms of their readiness for reading) for Model 1 reached 94% for those in Level 1 ("needs support"). The model reached 83% of true positives for students in Level 2 ("adequate"), and 74% of true positives for the predictive classification of students in Level 3 ("superior"). The level of overall accuracy for Model 1 was 84%. Results were highly successful for all classifications in general and in particular for the predictive classification of the category of highest interest, for which the neural network's architecture had been optimized.

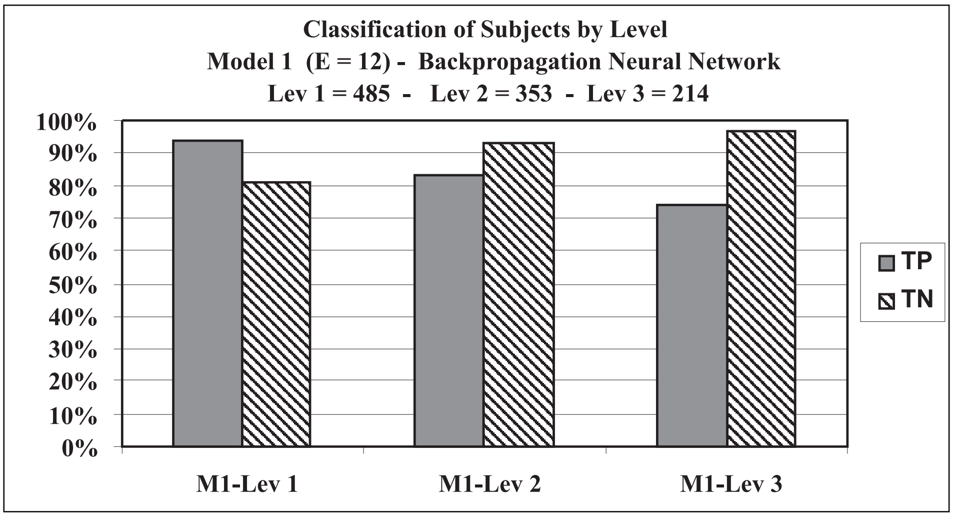

Figure 1. Outcome of Testing Phase for Model 1 (True Positive, True Negatives).

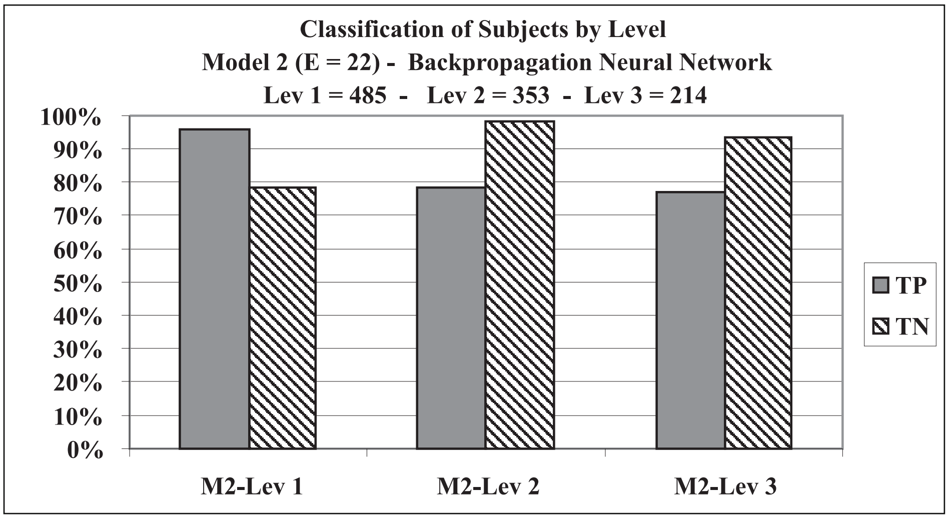

Figure 2. Outcome of Testing Phase for Model 2 (True Positive, True Negatives).

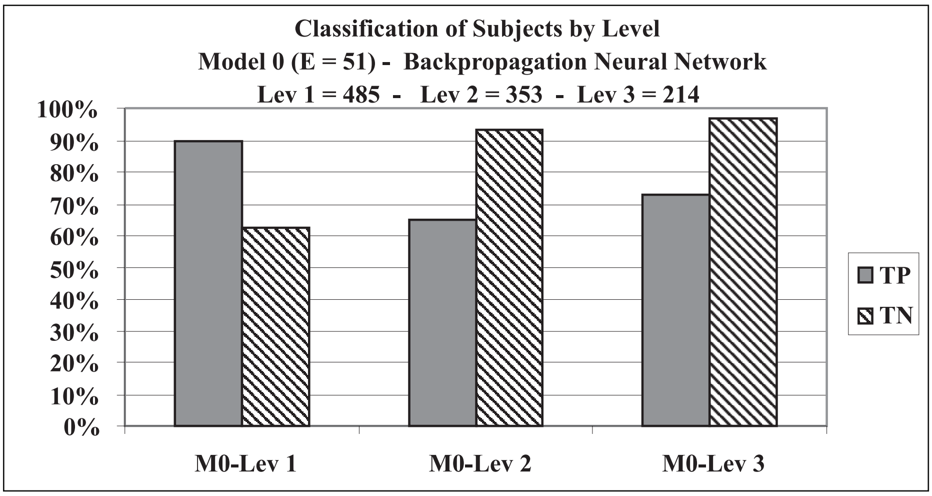

Figure 3. Outcome of Testing Phase for Model 0 (True Positive, True Negatives).

Further analyses

After obtaining the levels of accuracy desired for the predictive classification models, the next phase of the analysis involved the use of predictive systems to determine student characteristics in each of the performance levels. This was accomplished using self-organizing maps, Kohonen neural networks (Kohonen, 1977, 1982, 1988) to group students according to the characteristics of the configuration patterns present in the predictor variables (inputs to the network). Several groupings were explored, and in order to obtain the most significant groupings with a valid set of communalities, a hierarchical grouping technique was used (applying Ward's method), to reduce the number of groupings initially obtained with the Kohonen network approach (Fukushima, 1980). Thus, significant groups of students were identified, and further analyses will be carried out to describe and understand the underlying characteristics of these groups and their relationship with the performance levels identified for the students in each of the groups when the predictive neural network models were developed. This procedure will allow a more comprehensive understanding of the underpinnings present in the patterns detected in the predictive classification, further accounting for the expected performance differences, and will eventually facilitate the development of intervention programs aimed at the variables with the highest impact in the models predicting expected level of performance.

Study 2: Predictive classification of estimated future performance in a mathematics test using a neural networks approach.

There was a need to identify students that would be in high and low performing groups in mathematics performance, prior to further mathematics courses and testing, in order to provide any necessary tutoring support and achieve proper course assignments. In view of the complexity of the task and the characteristics of the model to be achieved, it was decided to that a neural network approach would be used to model the expected performance levels. This is the first analysis carried out with this methodology in the area of mathematics performance, and is used as a "proof of concept" of the significant potential of this approach to become a very powerful predictive assessment and evaluation methodology, based on existing data, that could eventually replace or complement traditional tests. The intrinsic power of the predictive system to capture the available patterns in the data and to relate them to expected outcomes could result in a significant advance for assessments in education and social sciences in general.

Objective: The specific purpose of the study then, was to develop predictive classification models that could identify with sufficient precision two groups of students corresponding to the highest 30% and lowest 30% of estimated future performance in a mathematics test, utilizing only cognitive, motivational and background variables, with no assessment of the mathematics content present in the test. It was expected that results would enable the development of an "early warning" system which would allow early and prompt intervention with those students most in need of support and remediation in mathematics (at the level of exit from secondary education, and beginning of university studies). Similarly, this approach could serve to identify top or advanced students and improve their placement and/or career choice. In addition, this research has strong implications in the understanding of cognitive, motivational and background variables that account for mathematics performance, and results in a proof-of-concept for this mode of developing predictive classifications for evaluative purposes in some educational assessment applications.

Methodology of Research (Study 2)

Subjects

The sample included 87 university students, of both genders (70% female), ages between 18 and 25, attending the first or second year of university studies in a large private university. Participating students belonged to the Psychology and Engineering departments.

Material and instruments

After giving informed consent, all the participants completed the following cognitive tasks, during the first session: 1) Attention Network Test (ANT) (Fan, McCandliss, Sommer, Raz, & Posner, 2002), and 2) Automated Operation Span (Unsworth, Heitz, Schrock & Engle, 2005). In a second session participants completed the Online Motivation Questionnaire (OMQ) (Boekaerts, 2002) and were given a mathematics test consisting of 65 items (50 items from a national mathematics test calibrated at exit from secondary education, and 15 items from the 12th grade TIMSS test administered world wide). In addition, an as in the previous study, a wide range of background variables was collected, including individual and family characteristics.

All stimuli of the cognitive tasks were presented via E-Prime software, on an IBM-compatible personal computer running Windows 2007, and presented on a 17-inch monitor, with a resolution of 1024 x 768. E- Prime is a software applications suite for conducting psychological and neuroscientific experiments, developed by Psychology Software Tools (PST). The distance between the subjects' eyes and the screen remained constant at 60 cm, for both cognitive tasks, maintaining the visual angle.

Attention Network Test (ANT) (Fan, McCandliss, Sommer, Raz, & Posner, 2002)**.** This task provides a measure for each of the three anatomically defined attentional networks: alerting, orienting, and executive. The ANT can be used as a phenotype of the efficiency of the attentional functions. In a small-scale twin study using the ANT, the executive network showed high-enough heritability (0.89) to justify the search for specific genes (Fan, Wu, Fossella, & Posner, 2001). The ANT is a combination of the cued reaction time (Posner, 1980) and the flanker test (Eriksen & Eriksen, 1974). Participants are asked to determine when a central arrow points left or right. The ANT's responses were collected via two mouse buttons (left-right). They were instructed to focus on a centrally located fixation cross throughout the task, and to respond as quickly and accurately as possible. During the practice trials, but not during the experimental trials, subjects received feedback from the computer on their speed and accuracy. The practice trials took approximately 2 min and each of the three experimental blocks took approximately 5 min. The whole experiment took about twenty minutes.

Stimuli consisted of a row of five visually presented horizontal black lines, with arrowheads pointing leftward or rightward, against a gray background (three flanker conditions). The target was a leftward or rightward arrowhead at the center. This target was flanked on either side by two arrows in the same direction (congruent condition), or in the opposite direction (incongruent condition), or by lines (neutral condition). The participants' task was to identify the direction of the centrally presented arrow by pressing the left mouse button for the left direction and the right mouse button for the right direction. A single arrow or line consisted of 0.558 of visual angle and the contours of adjacent arrows or lines were separated by 0.068 of visual angle. The stimuli (one central arrow plus four flankers) consisted of a total of 3.088 of visual angle. Each trial consisted of five events. First, there was a fixation period for a randomly variable duration (400- 1600 msec). Then, a warning cue was presented for 100 msec. There was a short fixation period for 400 msec after the warning cue and then the target and flankers appeared simultaneously. The target and flankers were presented until the participant responded, but for no longer than 1700 msec. After participants made a response, the target and flankers disappeared immediately and there was a post-target fixation period for a variable duration which was based on the duration of the first fixation and the reaction time (RT) (3500 msec minus the duration of the first fixation minus the RT). After this interval the next trial began. Each trial lasted for a total of 4000 msec. The fixation cross appeared at the center of the screen during the whole trial. To introduce an attentional-orienting component to the task, the row of five stimuli were presented in one of two locations outside the point at which the subject was fixating, at an angle either 1.068 above or below the fixation point. Target location was always uncertain except when spatial cues were presented.

To measure alerting and/or orienting attention, there were four warning conditions: no cue, center cue, double cue, and spatial cue. For the no-cue trials, participants saw only a fixation point for 100 msec. Under this condition, there were neither alerting nor spatial cues. For the center-cue trials, participants were shown an asterisk at the location of the fixation cross for 100 msec. Therefore, alerting was involved. For the double-cue trials, the time course was the same as in the center cue trials except that there were two warning cues corresponding to the two possible target positions—up and down. It was expected that the alerting response was involved but the attentional field was larger under the double-cue condition than under the central-cue condition. For the spatial-cue trials, the cue was at the target position and the time course was the same as in the center-cue and double-cue trials. The spatial cues were always valid, which means that they were displayed right on the locations of the targets. It was expected that both alerting and orienting were involved under this condition. The variable duration of the first fixation was used to produce additional uncertainty about cue onset (Fan, McCandliss, Sommer, Raz, & Posner, 2002). The efficiency of the three attentional networks is assessed by measuring how response times are influenced by alerting cues, spatial cues, and flankers (Fan et al., 2002).

Automated Ospan. This is a computer-administered version of the Ospan instrument (Unsworth, Heitz, Schrock & Engle, 2005) that measures working memory capacity. The responses were collected via click of a mouse button. First, the subject responded to three practice sessions. In the first practice block, letters appeared sequentially on the screen, and the participants were required to recall the letters in the same order in which they were presented. In all experimental conditions letters remained on the screen for 800 msec. At recall, the participants saw a 4 x 3 matrix of letters (F, H, J, K, L, N, P, Q, R, S, T, and Y). Recall consisted of clicking the box next to the appropriate letters (no verbal response was required) in the correct order. The recall phase was untimed. After recall, the computer provided feedback about the number of letters correctly recalled in the current set. In the second practice block, the participants practiced the math exercises of the task. They first saw a math operation (e.g., (1*2) + 1=?). The participants were instructed to solve the operation as quickly and accurately as possible and then click the mouse to advance to the next screen. On the next screen a digit (e.g., 3) was presented and the participants were required to click either a "true" or "false" box, depending on their answer to the previously shown math problem. After each operation, the participants were given accuracy feedback. The math practice served to familiarize them with the math exercises of the task as well as to calculate the reaction time to solve math operations. After the math practice block, the program calculated each individual's mean time required to solve the equations. This time (plus 2.5 SD) was then used as a time limit for the math exercises of the experimental session for that individual. The participants completed 15 math operations in the practice block. In the final practice block, the participants performed both the letter recall and math portions together, just as they would do in the experimental blocks trials. The participants first saw the math operation, and after they clicked the mouse button indicating that they had solved it, they saw the letter to be recalled. If the participants took more time to solve the math operations than their average time plus 2.5 SD, the program automatically moved on and counted that trial as an error. Participants completed three practice trials each of set size 2. After participants completed all of the practice blocks, the program progressed to the experimental trials, which consisted of three sets of each set size, with the set sizes ranging from 3 to 7. This made for a total of 75 letter- and 75 math-problems. The order of set sizes was random for each participant. An 85% accuracy criterion was imposed for all participants. Therefore, they were encouraged to keep their math accuracy at or above 85% at all times. During recall, a percentage in red was presented in the upper right-hand corner of the screen, indicating the percentage of correctly solved math operations (Unsworth, Heitz, Schrock & Engle, 2005). The program reported five scores: Ospan score, total number correct, math errors, speed errors, and accuracy errors. Ospan score was the sum of all perfectly recalled sets. The "total number correct" score was the total number of letters recalled in the correct position across all trials. Three types of errors were reported: "Math errors" were the total number of task errors, which was then broken down into "speed errors," in which the participant ran out of time in attempting to solve a given math operation, and "accuracy errors," in which the participant solved the math operation incorrectly. The task took approximately 20–25 minutes to complete (Unsworth, Heitz, Schrock & Engle, 2005).

On-line Motivation Questionnaire. The last version of the On-line Motivation Questionnaire, namely the OMQ91 (Boekaerts, 2002) was used to study motivational variables. This is a self-report questionnaire that consists of two parts (Part 1: before the task; Part 2: after the task). Part 1 (appraisal part) included 23 items that measure three aspects of task motivation: appraisals, emotions, and learning intention. Appraisals are registered in 13 items that measure three aspects of task judgment: personal relevance of the curricular task (e.g. How useful do you consider this task), subjective competence (e.g. How good are you at doing this type of task?), and task attraction (e.g. How much do you like this type of task?). Six items refer to emotional state (e.g. How do you feel now? Nervous…not nervous; happy…not happy). Four items measure learning intention (e. g. How much effort are you going to invest in the task) Part 2 of the OMQ (Attributions) consists of 23 items of which 7 measure the emotion students felt after doing the task. Three items measure invested effort (e. g. "How hard did you work on this task?"), 2 items register result assessment (e. g. "How well did you do on this task?), and 8 items measure the attribution process. Students are requested to select either the questions that relate to successful or unsuccessful task completion (e .g. "I completed this task (un)successfully, because I am (not) competent to do this type of task).

Mathematics Test. This test consisted of 65 multiple choice items with four or five options and only one correct answer (50 items were taken from a national test (Cortada de Kohan & Macbeth, 2007) and 15 items were extracted from disclosed items of the Trends in International Mathematics and Science Study (TIMSS, 1995), all calibrated by 3-parameter IRT analysis). The items measure simple algorithms for arithmetic problems: some items required the use of percentages or proportions, decimal numbers, and a few others are algebraic and geometric questions. There was no time limit to take the test, but its duration for all students was under two hours.

Background information: basic background information of each student, family system, socio-economic data, level of education, occupation, and other similar variables.

Model

The procedure used was again a back propagation network, that is, a multilayer network composed of nonlinear units, which computes its activation level by summing all the weighted activations it receives and which then transforms its activation into a response via a nonlinear transfer function. All other characteristics of the model were similar to the ones described in the previous study.

Architecture of the neural networks

Two different neural networks (NN) were developed as predictive systems for the two tasks of this study. NN1 was developed to maximize the predictive classification of the highest 30% of students, which would be scoring the highest in the mathematics test. NN2 was developed to maximize the predictive classification of the lowest 30% of students, which would be scoring the lowest in the mathematics test. The specific architecture of each of the two neural networks developed is as follows:

NN1: All cognitive, motivational, and background variables were introduced in the analysis. They were used for the development of the vector-matrix containing all predictor variables for each student. The resulting network contained all the input predictors, some of them collapsed into subscales to maximize predictive classification, with a total of 36 input units. The model built contained one hidden layer, with 10 units, and an output layer with two units (categories corresponding to "belongs to top 30%" or "belongs to lower 70 %"). A standardized method for the rescaling of covariates was used. The hidden layer had a hyperbolic tangent activation function. For the output layer, the activation function chosen was identity, and the error function the sum of squares.

NN2: All cognitive, motivational, and background variables were introduced in the analysis. They were used for the development of the vector-matrix containing all predictor variables for each student. The resulting network contained all the input predictors, some of them collapsed into subscales to maximize predictive classification, with a total of 37 input units. The model built contained two hidden layers, with 10 and 8 units, respectively. The output layer contained two units (categories corresponding to "belongs to Lowest 30%" or "belongs to Highest 70 %"). A standardized method for the rescaling of covariates was used. The hidden layers had hyperbolic tangent activation functions. For the output layer, the activation function chosen was also hyperbolic tangent, and the error function the sum of squares.

Procedure

The software used was SPSS v.16 – Neural Network Module, for the development and analysis of all predictive models in this study. The usual three development phases of the predictive system were carried out: training of the network, testing of the network developed, and validation of the network. During the training phase several models were attempted, and several modifications of the neural network parameters were tried, manipulating learning persistence, learning rate, momentum, and other criteria. These tests continued until achieving desired levels of classification, maximizing the benefits of the model chosen. In this analysis both precision and recall, as outcome measures of the network, were given equal weight. There was no need to trim the number of predictor inputs into the model, both for NN1 and NN2.

Results of Research (Study 2)

The results obtained were very successful in terms of the effectiveness of the two predictive classificatory models developed. After various iterations and testing of alternative architectures the two final models were developed one each for the two predicted classification desired. Both models were able to reach 100% correct identification of all students, both in the target group for each network (Highest 30% of scores or Lowest 30% of scores, respectively) and those "not belonging" to the target group (see Tables 2 and 3). The model required for the classification of the lowest scores was somewhat more complex, and required two hidden layers in the architecture of the neural network.

Table 2. Predictive classification of performance levels in mathematics test, modeling with basic cognitive processing, motivational and background variables, for the predictive identification of the Highest 30% of math scores.

Testing phase of neural network (NN1 – 30% Highest group)			
		Predicted Performance	
		<30% Highest	30% Highest
Observed Performance	< 30% Highest	100%	0%
	30% Highest	0%	100%

Table 3. Predictive classification of performance levels in mathematics test, modeling with basic cognitive processing, motivational and background variables, for the predictive identification of the Lowest 30% of math scores.

Testing phase of neural network (NN2 – 30% Lowest group)			
		Predicted Performance	
		<30% Lowest	30% Lowest
Observed Performance	< 30% Lowest	100%	0%
	30% Lowest	0%	100%

In addition, both networks showed interesting differences in the pattern of relative normalized importance of those variables with the highest participation in the predictive model. For the Low performers (those predicted to be in the lowest 30% of scores), several basic cognitive variables were most important, such as "reaction-time", "working memory capacity", and the closely related "executive attention", all having to do with the control and the speed of processing.. In fact, three out of the top four variables in terms of relative predictive importance correspond to basic cognitive processing variables (see Table 4), with high relative values. Among the self-regulation variables, only "expected results of the assessment" appeared among the most predictive.

Table 4. Relative importance of the top variables participating in the model for the predictive classification of the Lowest 30% of scores in the mathematics test.

Independent Variable Importance – Low 30% Group		
	Importance	**Normalized Importance %**
Gender	0.035	34.2
Mother's educational level	0.028	28.2
Father's educational level	0.024	23.9
Mother's occupation	0.065	64.5

Father's occupation	0.059	58.8
Age	0.062	61.5
Competence-related attribution for success	0.041	40.3
Personal Relevance of Task	0.029	28.3
Subjective Competence	0.043	42.7
Task Attraction	0.042	41.8
Learning Intention	0.052	51.8
Reported Effort	0.062	61.1
Expected Result of Assessment	0.099	97.7
Emotional State	0.062	61.3
Alerting Attention	0.029	29.1
Orienting Attention	0.018	17.4
Executive Attention	0.067	66.4
Working Memory	0.081	80.6
Reaction Time (operations)	0.101	100.0

On the other hand, results from the predictive model for those expected to be in the Highest 30% of the scores, the top three predictors with the most significant participation were "task attraction", "father's occupation", and "reported effort", all among the self-regulation and background variables. Only "working memory capacity" (as measured by "absolute AOSPAN") among the basic cognitive processing variables, appeared among the top five predictors, and then with a much lower relative importance than for the Low 30% group (see Table 5). It is quite evident the relative lower importance of all cognitive control and speed of processing variables, which do not discriminating well for the predictive classification in the Highest 30% group (see Table 5). is also worth noting the relative high importance of parent's occupation in both Low and High group, particularly in the latter.

Table 5. Relative importance of the top variables participating in the model for the predictive classification of the Highest 30% of score in the mathematics test.

Independent Variable Importance – High 30% Group		
	Importance	Normalized Importance %
Mother's educational level	0.041	31.5
Father's educational level	0.055	42.4
Mother's occupation	0.060	46.5
Father's occupation	0.086	66.4
Age	0.044	34.0
Competence-related attribution for success	0.046	35.5
Personal Relevance of Task	0.033	25.2
Subjective Competence	0.058	45.0
Task Attraction	0.130	100.0
Learning Intention	0.044	33.6
Reported Effort	0.078	59.9

Expected Result of Assessment	0.028	21.6
Emotional State	0.052	40.3
Alerting Attention	0.036	27.9
Orienting Attention	0.022	17.1
Executive Attention	0.050	38.8
Working Memory	0.077	59.5
Reaction Time (operations)	0.060	46.2

It is clear from these results that besides the evident predictive power of both neural networks to model the expected performance of both the low and high performance groups, this methodology has also detected important differences in the factors that seem to underlie the students' performance. Among the lowest 30% of the student group, the main determinants of performance appear to be the basic cognitive processing variables, indicating the degree to which they represent the areas of relative weakness in the group, and more discriminating from the rest of the students. On the other hand, among the high 30% of the student group, the main determinants appear to be self-regulation and background variables (particularly how interested they were in the task, and social indicators such as parents' occupations), leaving the cognitive variables at much lower levels of importance, probably also due to the fact that they are more evenly represented in the group and are therefore less discriminating.

Discussion

These highly successful examples of the use of a predictive systems approach in the field of educational assessment show that it is a methodology that can make a very useful contribution to the field. Because of the complexity of the variables involved in the field of education and educational assessment in particular, it is possible to take full strengths of a predictive systems approach in order to improve the quality of educational assessments. In particular, neural networks, with their superior potential for pattern recognition and classification are particularly well suited to use available data to give a much broader expectation of academic performance, based on a wider set of predictors which lead to an increase validity of the construct and the obtained results, while at the same time increasing the accuracy of the resulting classifications.

Increasing the validity also makes the results more generalizable, and as it has been pointed out, the generalizability of an assessment outcome is a crucial aspect as criteria for establishing the quality of an assessment (Messick, 1995; Gielen, Dochy & Dierick, 2003). Neural network approaches offer a strong method to achieve these results.

The predictive systems approach allows for the understanding of the students' individual characteristics, in addition to the prediction of expected performance levels. This opens major possibilities for improvement of evaluation procedures and the planning of interventions. In addition it has implications for the application of these methods in educational research and in the implementation of diagnostic "early-warning" programs in educational settings, as well as informing cognitive theory and the development of automated tutoring and learning systems. The capacity to very accurately classify students, which is also what tests attempt to do (usually with much less success in terms of accurate classifications), without the performance sampling issues of traditional testing is a major step in using a much broader picture of a student's overall inputs into any attempted performance, and therefore a much more valid approach to educational evaluation. In turn, this new approach allows for the conceptualization and development of new modes of assessment which could facilitate breaking away from traditional forms of testing while at the same time improving the quality of the assessment (Segers, Dochy & Cascallar, 2003). Such new methods could achieve greater precision and accuracy, while including more of the effects that intervene significantly in the determination of specific performances as well as overall academic performance, and are conducive to the understanding of academic performances in a broader and better specified theoretical framework, with the support of a statistical and mathematical model that can deal satisfactorily with such richness and complexity of information. These methods will also be able to detect more effectively subtle patterns and interactions in the universe of complex educational data with greater reliability and accuracy, which can not be matched by traditional statistical approaches.

Conclusions

A predictive systems approach, and the resulting operational models, together with the conceptual developments previously mentioned in the areas of methodology and assessment, are of fundamental importance for the development of truly new assessment instruments, programs for the assessment of various academic performances, and for diagnostic and selection purposes in the educational field. Thus, the role of educational assessment can go through a radical shift, conceptually unifying the areas of formative and summative assessment, with broader hybrid models which integrate, with greater validity, accuracy, and utility, different goals which are at once evaluative and predictive. Eventually, these new directions can lead to better assessment programs, improved diagnostic and placement evaluations, better admission systems, and "continuous assessment" in the context of "intelligent classrooms" engineered to capture all outputs of students in real time, monitoring their performance, and by linking it to exisiting data bases and statistical anchors, they could provide continuous, ongoing evaluation of students' progress and level of performance. These possible applications, although somewhat in the future, are certainly closer to being achieved with the introduction of predictive systems approaches in educational research and assessment.

References

Al-Deek, H. M. (2001). Which method is better for developing freight planning models at seaports – Neural networks or multiple regression? *Transportation research record,* 1763, 90–97.

Bansal, A., Kauffman, R., & Weitz, R. (1993). Comparing the Modeling Performance of Regression and Neural Networks as Data Quality Varies: A Business Value Approach. *Journal of Management Information Systems,* 10*,* 11–32.

Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in Motivation and Achievement. New Directions in Measures and Methods*, (Vol. 12, pp. 77–120). New York, JAI: Elsevier Science.

Boekaerts, M., & Cascallar, E. (2006). How far have we moved toward the integration of theory and practice in self-regulation? *Educational Psychology Review*, 18(3), 199–210.

Braun, H. J. (2005). *Using student progress to evaluate teachers: A primer on value added models*. New Jersey: Educational Testing Service.

Cascallar, E. C. (2003). *Applications of IRT for pruning of inputs in predictive systems*. (Technical Report). Washington, DC: American Institutes for Research.

Cascallar, E. C., Boekaerts, M., & Costigan, T. E. (2006) Assessment in the Evaluation of Self-Regulation as a Process, *Educational Psychology Review*, 18(3), 297–306.

Cascallar, E. C. & Musso, M. (2008). Classificatory Stream Analysis in the Prediction of Expected Reading Readiness: Understanding Student Performance. *International Journal of Psychology, XXIX International Congress of Psychology ICP 2008*, 43, (3/4): 231.

Cortada de Kohan, N. & Macbeth, G. (2007). Construcción de un test de matemática para adolescentes y adultos. *Interdisciplinaria,* 24(1), 43–64.

Detienne, K. B., Detienne D. H., & Joshi, S. A. (2003). Neural Networks as Statistical Tools for Business Researchers. *Organizational Research Methods,* 6, 236–265. Downloaded from http://orm.sagepub.com on October 20, 2008.

Duliba, K. (1991). *Contrasting neural nets with regression in predicting performance*. Proceedings of the 24th Hawaii International Conference on System Sciences, 4, 163–170.

European Commission – Directorate-General for Education and Culture (2000). *European report on the quality of school education*. Brussels.

Eriksen, B. A. & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a non search task. *Psychophysics,* 16(1), 143–149.

Everson, H. T. (1995). Modeling the student in intelligent tutoring systems: The promise of a new psychometrics. *Instructional Science,* 23 (5–6), 433–452.

Everson, H. T., Chance, D., & Lykins, S. (1994). *Exploring the use of artificial neural networks in educational research.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Fan, J., Wu, Y., Fossella, J., & Posner, M. I. (2001). Assessing the heritability of attentional networks. *BMC Neuroscience,* 2, 14–20.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience,* 14 (3), 340–347.

Fukushima, F. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics,* 36(4), 193–202.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessments effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimizing New Modes of Assessment: In Search of Qualities and Standards,* (pp. 37–54). The Netherlands: Kluwer Academic Publishers.

Gorr, W. (1994). Research Perspective on Neural Networks. *International Journal of Forecasting*, 10(1), 1–4.

Grossberg, S. (1980). How Does the Brain Build a Cognitive Code?. *Psychological Review,* 87, 1–51. Reprinted in 1988, J. Anderson & E. Rosenfeld, Eds., *Neurocomputing: Foundations of Research,* Cambridge, MA, MIT Press.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control.* Boston: Reidel Press.

Hardgrave, B. C., Wilson, R. L., & Walstrom (1994). Predicting graduate student success: A comparison of neural networks and traditional techniques. *Computers and Operations Research,* 21(3), 249–263.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd edition). New York: Prentice Hall.

Kohonen, T. (1977). *Associative Memory. A system theoretical approach*. Berlin: Springer-Verlag.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69. Reprinted in 1988, J. Anderson & E. Rosenfeld, Eds., Neurocomputing*: Foundations of Research,* Cambridge, MA, MIT Press.

Kohonen, T. (1988). An introduction to Neural Computing. *Neural Networks,* 1, 3–16.

Laguna, M. & Marti, R. (2002). Neural network prediction in a system for optimizing simulations. *IIE Transactions,* 34 (3), 273–282.

Lippman, R. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine,* 3(4), 4–22.

Marquez, L., Hill, T., Worthley, R., & Remus, W. (1991). Neural network models as an alternative to regression. *Proceedings of the IEEE 24th Annual Hawaii International Conference on Systems Sciences,* 4, 129–135.

Mavrovouniotis, M. L. & Chang, S. (1992). Hierarchical neural networks. *Computers & Chemical Engineering,* 16(4), 347–369.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), 741–749.

Neal, W. & Wurst, J. (2001). Advances in Market Segmentation. *Marketing Research*, 13(1), 14–18.

Nguyen, N. & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression and artificial neural networks. *Journal of Real Estate Research,* 22(3), 313–336.

Perkins, K., Gupta, L. & Tammana (1995). Predict item difficulty in a reading comprehension test with an artificial neural network. *Language Testing,* 12(1), 34–53.

Posner, M. I. (1980). Orientation of attention: The VIIth Sir Frederic Bartlett lecture. *Quarterly Journal of Experimental Psychology,* 32, 3–25.

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Volume I. Cambridge, MA: MIT Press.

Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning representations by back-propagating errors. *Nature,* 323, 533–536.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing new modes of assessment: in Search of Qualities and Standards.* The Netherlands: Kluwer Academic Publishers.

Specht, D. (1991). A general regression neural network. *IEEE Transactions on Neural Networks,* 2(6), 568–576.

Sternberg, R. J. (2000). The concept of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 3–15). New York: Cambridge University Press.

Trends in international Mathematics and Science Study (TIMSS) (1995). Prepared by the International Association for the Evaluation of Educational Achievement (IEA). Amsterdam. Downloaded from http://timss.bc.edu/timss1995i/TIMSSPDF/C_items.pdf on March, 2008.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods,* 37 (3), 498–505.

Weiss, S. M. & Kulikowski, C. A. (1991). *Computer systems that learn*. San Mateo, CA: Morgan Kaufmann Publishers.

White, H. & Racine, J. (2001). Statistical inference, the bootstrap, and neural-networking modeling with application to foreign exchange rates. *IEEE Transactions on Neural networks,* 12(4), 657–673.

Wilson, R. L. & Hardgrave, B. C. (1995). Predicting graduate student success in a MBA program: Regression vs. Classification. *Educational and Psychological Measurement,* 55(2), 186–195.

Adviced by Zarko Vukmirovic, American Institutes for Research, Washington, USA

Mariel Musso Ph.D., Researcher, Leiden University, Netherlands; Assessment Group International, USA.
Bld. Louis Schmidt 2C (bte 20), 1040 – Brussels, Belgium.
Phone: +32.474.244.244.
E-mail: agi_group@msn.com
Website: http://www.leidenuniv.nl/

Eduardo Cascallar Ph.D., Professor, Leiden University, Netherlands; Catholic University of Leuven, Belgium; Assessment Group International, USA.
Email: cascallar@msn.com
Website: http://www.kuleuven.be/english/