Journal of Autism and Developmental Disorders, Vol. 23, No. 3, 1993

A Neural Network Approach to the Classification of Autism[1]

Ira L. Cohen[2] and Vicki Sudhalter

New York State Office of Mental Retardation and Developmental Disabilities, New York State Institute for Basic Research in Developmental Disabilities, Staten Island

Donna Landon-Jimenez

Queens College and the Graduate Center of the City University of New York and CSI/IBR Center for Developmental Neuroscience

Maryellen Keogh

New York State Office of Mental Retardation and Developmental Disabilities, New York State Institute for Basic Research in Developmental Disabilities, Staten Island

A nonlinear pattern recognition system, neural network technology, was explored for its utility in assisting in the classification of autism. It was compared with a more traditional approach, simultaneous and stepwise linear discriminant analyses, in terms of the ability of each methodology to both classify and predict persons as having autism or mental retardation based on information obtained from a new structured parent interview: the Autistic Behavior Interview. The neural network methodology was superior to discriminant function analysis both in its ability to classify groups (92 vs. 85%) and to generalize *to* new *cases that were not part of the training sample (92*

[1] The authors are gratefully indebted to Dov J. Shazeer, Group Leader, Image Recognition Systems, Charles Stark Draper Laboratory, Cambridge, Massachusetts, for initially suggesting the use of neural networks for problems in diagnosis and for his considerate guidance in helping the first author to understand the complexity of this methodology. In addition, we are also indebted to the staff at NeuralWare, Inc., Pittsburgh, Pennsylvania, in particular, Bob Everly, for his kind advice, patience, and assistance to the first author and to Allan Reiss, and Enid G. Wolf-Schein, for their comments on the Autistic Behavior Interview. This work was supported by funds from the New York State Office of Mental Retardation and Developmental Disabilities.

[2] Address all correspondence to Ira L. Cohen, Department of Psychology, New York State Institute for Basic Research in Developmental Disabilities, 1050 Forest Hill Road, Staten Island, New York 10314.

vs. 82%). Interrater and test–retest reliabilities and measures of internal consistency were satisfactory for most of the subscales in the Autistic Behavior Interview. The implications of neural network technology for diagnosis, in general, and for understanding of possible core deficits in autism are discussed.

Diagnosing autism presents a difficult problem for researchers and clinicians alike since the classification criteria and the precision of their definitions have varied over the years. Autism was first recognized as a unique disorder by Kanner (1943) in his seminal paper, entitled "Autistic Disturbances of Affective Contact," which provided very rich narrative descriptions of the condition. Subsequent generations of researchers and clinicians have attempted to provide various criteria for diagnosis of this condition such as that provided by the British Working Party (Creak, 1961), Ritvo and Freeman (1978), Rutter (1978), DSM-III (American Psychiatric Association ([APA]), 1980), and DSM-III-R (APA, 1987) with each emphasizing different behavioral criteria.

The most recent changes provided in the DSM-III-R have made autism easier to define, especially in adults. However, it is still unclear whether these criteria define the same entity that was identified as autism using earlier standards (Volkmar, Bregman, Cohen, & Cicchetti, 1988). Additionally, even with the DSM-III-R criteria, the decision to classify an individual as autistic is obviously influenced by many other factors including the experience of the diagnostician, his or her training and personal beliefs. For example, many severely impaired, developmentally disabled individuals display behaviors that might qualify them as autistic to some diagnosticians. To others, the behaviors are categorized as "autistic features" and attributed to the person's retardation; a phenomenon that could reflect "diagnostic overshadowing" (Reiss, Levitan, & Szyszko, 1982) if the diagnosis of autism is, in fact, appropriate.

These problems in identifying "signals" of autism and establishing objective cutoff criteria clearly have an impact on clinicians who must make appropriate treatment recommendations as well as on researchers. Clinically, many persons with autism are often not diagnosed as such because there are not enough diagnosticians around who are knowledgeable about the variety of ways autism can express itself. This is not a trivial issue since diagnosis has a strong influence on right to effective treatment, placement, treatment contraindications, and so forth. For researchers attempting to understand autism, the frequent changes in criteria through the years may indicate that these scientists have not always been studying the same classes of children, even though their labels were the same. Thus, what we think we know about autism may only be true of subsets of children or subsets of behaviors that consistently appear in the various diagnostic criteria.

Researchers have attempted to surmount issues regarding reliability and specificity of diagnostic criteria by devising various rating scales to measure the behaviors associated with autism and then validating these against an independent diagnostic judgment. However, there is no single behaviorally based diagnostic instrument accepted or used by clinicians or researchers (Parks, 1983). To cite Rapin (1987): "Experienced clinicians rarely rely on checklists for diagnosing autism. They find it easy to spot an autistic child and to reach consensus" (p. 712). Further, while the ability of such instruments to classify persons with autism from those with other disorders has been assessed in a number of studies using multivariate techniques (e.g., Wadden, Bryson, & Rodger, 1991), the ability of these multivariate models to then predict classification of a *different* data set with a high degree of accuracy remains to be determined. Although experienced clinicians may intuitively recognize the "core behavioral syndrome of autism" (Rapin, 1987; p. 712) as a gestalt and therefore not need rating scales, defining and measuring the core symptomatology has remained an elusive task. There appear to be at least several reasons for this.

First, researchers do not yet know or agree on the definition of the "core deficit(s)," of autism, yet alone on how to best measure it (them) with some emphasizing social deficits (Wing & Attwood, 1987), others emphasizing language (Rutter, Bartak, & Newman, 1971), and still others sensory/perceptual disturbance (Ornitz & Ritvo, 1968).

Second, our measures of behaviors associated with autism are inherently "noisy." In a study of the reliability of several autism assessment instruments, Sevin, Matson, Coe, Fee, and Sevin (1991) found Pearson interrater reliability correlations ranging from .32 to .89 across items for the Real Life Rating Scale (Freeman et al., 1986) and .14 to .85 across items for the Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1988). Kappa coefficients for the Autism Diagnostic Interview (Le Couteur et al., 1989) range from .64 to .97 across items within subscales for trained interviewers.

Third, currently accepted diagnostic systems use a logical decision-making algorithm such as that used in the DSM-III-R. However, as noted above, interpretational problems may arise in those cases in which this algorithm classifies persons as autistic when other methods of classification (e.g., DSM-III), do not (Volkmar et al., 1988). More to the point, it has not been demonstrated that the diagnostic process is as logical as the DSMs imply.

Finally, researchers have attempted to identify or verify the validity of symptom constructs by using linear discriminant analysis to separate predefined groups (e.g., Krug, Arick, & Almond, 1980; Wadden et al., 1991). However, certain behavior patterns thought to be uniquely associated with autism may not, *in principle*, be validated as such using this and other similar multivariate procedures as they are traditionally used. Typical discriminant analysis procedures assume that the groups that are to be separated and defined as distinct from each other can be divided by a line, plane, or hyperplane, depending on the number of dimensions to be separated. That is, it is assumed that a linear solution is available. But even some very simple problems have outcomes that cannot be divided easily in this manner, no matter how many variables are examined because the outcome results from a nonlinear function (Rumelhart, Hinton, & Williams, 1986).

For example, let us assume that a behavioral checklist provided two descriptors of dysfunctions identified with autism: (a) gaze avoidance and (b) topic perseveration. Let us further assume that autism would be expected if an older, verbal person exhibited either of these behaviors, but not both. Let us suggest, instead, that this combination is indicative of social anxiety rather than autism per se. The absence of both would likely indicate, therefore, that aberrant social interaction was not a problem and so autism would be highly unlikely. This logical "exclusive OR" outcome is plotted in a two-dimensional coordinate system in Figure 1. Note that while the presence of either, but not both, descriptors is indicative of autism, there is no way to draw a single line that will isolate this pattern from the other possibilities and so this pattern of deviance in social communicative interaction could not be identified as uniquely associated with autism using traditional linear classification techniques. Instead, with two dimensions, two lines or a closed curve would be required to isolate this pattern. It is true that adding a third dimension, the multiplicative interaction of both descriptors, could be used to solve this problem. However, such interactions are almost never used in traditional discriminant analyses. Even if they were, however, the interactions would still be linear in nature and straight lines cannot be used to approximate nonlinear functions, should these functions provide the best descriptions of the data (e.g., sine waves, but not straight lines, approximate complex functions in Fourier analysis). As Wing and Attwood (1987) noted in discussing differential diagnosis of autism, "nature never draws a straight line without smudging it" (p. 12).

We argue that the diagnostic process is not one that can be mimicked easily by a logical decision-making algorithm but, instead, is more akin to the processes involved in pattern recognition, that is, we are arguing for a "right" as opposed to a "left" hemisphere approach to this problem. As such, solving the diagnostic dilemma of autism calls for the use of nonlinear pattern recognition techniques.

A new methodology has been developed recently which has features that make it especially attractive for such problems. Computer programs have been developed that mimic, at a very primitive level, the massively parallel processing pattern recognition abilities of the nervous system. Such "neural networks" have the following properties (Rumelhart & McClelland, 1986; NeuralWare, 1991) pertinent to the problems raised above:

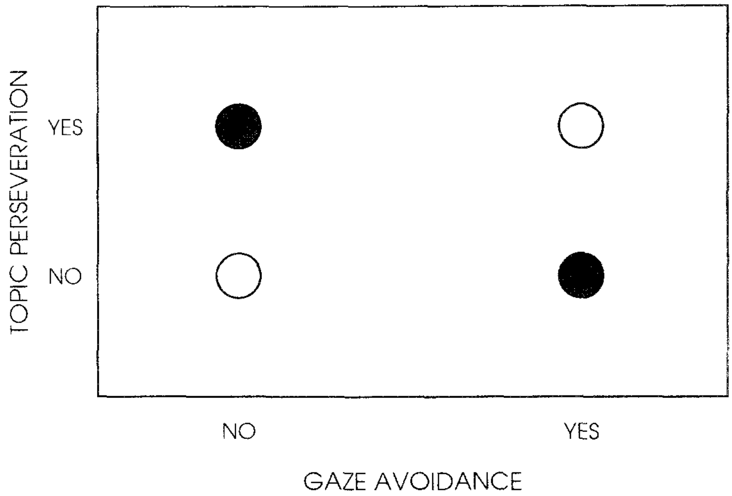

Fig. 1. A diagram of the "exclusive OR" problem as applied to diagnosis of autism. In this hypothetical situation, a diagnosis of autism is appropriate if either of the behaviors of gaze avoidance or topic perseveration is present (filled circles) but not if neither is present or both are present (open circles). Note that these two different patterns cannot be separated by a single line, as would be performed in a two-group discriminant analysis.

1. Unlike "expert system" diagnostic models and currently used methods for arriving at a diagnosis that rely on a priori logically defined rules (e.g., two items from section A AND 1 each from section B AND C AND a total number of items greater than or equal to 8), neural networks form their own "concepts" by learning from examples (training), much the same as diagnosticians learn what autism is by being shown a variety of cases by their mentors. For example, Sejnowski and Rosenberg have taught a neural network to read. In the process of learning, it formed several concepts such as recognition of vowels and consonants, spaces between words, and so forth (cited in NeuralWare, 1991). Other networks have been taught to read handwriting, a problem that cannot be readily solved using a linear approach (NeuralWare, 1991). Surely, the process by which a diagnostician arrives at a diagnosis of autism must, at least, be as complicated as the process by which that person deciphers another diagnostician's handwriting. If network models of diagnosis can be found that are highly accurate, then the concepts they infer could, perhaps, prove helpful in defining the observable core characteristics of autism.

2. Stored information (memory) in such programs is "distributed" throughout the network. This means that a trained network is capable of generating an outcome even if the input is noisy, incomplete, missing, or has never been presented to the network during training. In other words, neural networks are capable of generalizing from previous experience. In fact, addition of noise to a network during training actually facilitates generalization by preventing the network from "memorizing" the data set if the data set is small (R. Everly, personal communication).

3. Discriminant analysis and other multivariate procedures assume that the data are normally distributed, a very restrictive assumption. Many neural networks make no assumptions about the underlying distribution of the data on which they make associations and are therefore nonparametric. This makes them excellent for the typical types of data collected with rating scales.

4. Information is processed in a complex, nonlinear fashion by the network. This often makes such networks much better categorizers than traditional linear classifiers (Rumelhart, Hinton, & McClelland, 1986) such as discriminant function analysis.

As a result, a variety of different neural network models have been found to be excellent pattern recognizers and have found use in language processing (Rumelhart & McClelland, 1986), data compression, signal processing, and medical diagnosis (cited in NeuralWare, 1991). Goodman et al. (1992) found a network employing "fuzzy logic" to be a better predictor of survival time after coronary artery bypass surgery than discriminant analysis and Weinstein et al. (1992) reported the superiority of a back-propagation network (see below) to discriminant analysis in elucidating and predicting the mechanism of action of cancer drugs. Others have used networks to model pathological processes (Grossberg, 1984).

The purpose of this study was to examine the relative utility of one type of neural network model, a feed-forward network using a back-propagation training algorithm, for classification of persons diagnosed with autistic disorder versus a matched group of persons with mental retardation who did not have autism. Diagnoses were based on clinical experience as well as the DSM-III (APA, 1980) and DSM-III-R (APA, 1987) criteria. The accuracy of diagnosis was compared with a more traditional means of classification—discriminant function analysis. Information about these persons was obtained by parent or caregiver interview using the Autistic Behavior Interview (ABI), an assessment instrument we developed recently. We felt, based on other rating scales such as the Vineland Adaptive Behavior Scales (VABS; Sparrow, Balla, & Cicchetti, 1984), that parent interviews could provide valuable information concerning the variety of behaviors displayed by persons with autism and that the interview technique is useful in explaining behavioral descriptors to persons who are likely to be less sophisticated in this matter than the interviewer (cf. Le Couteur et al., 1989).

Interrater and test–retest reliabilities and internal consistency measures were obtained on a subset of the data.

METHOD

Subjects

The parents or caregivers of 138 persons were interviewed, 69 of whom met the DSM-III (APA, 1980) criteria for Infantile Autism and the DSM-III-R (APA, 1987) criteria for Autistic Disorder based on observations, record review, and informal interviews and clinical judgment (I.L.C. and V.S.). The other 69 cases consisted of persons with mental retardation who did not meet these same diagnostic criteria.

The two groups were matched on chronological age and level of motor skills as defined by the Motor Skills standard score on the VABS. Groups were matched on motor skills since impairment in this area is not typically associated with autism whereas problems in communication and socialization, two other measures on the VABS, are clearly affected.

The mean ages (*SD*) of the autistic and nonautistic groups were 9.57 (7.08) and 8.45 (5.33) years, respectively, $t(136) = -1.04, p = .30$. The mean (*SD*) standard motor skills scores were 64.96 (21.66) and 71.06 (21.90), respectively, $t(136) = 1.63, p = .11$. Eight of the persons with autism and 13 of the persons with retardation were female. This difference was not statistically significant ($\chi^2 = 1.40$, $p = .24$). Of the mentally retarded cases, 19 also met the criteria for Pervasive Developmental Disorder Not Otherwise Specified (PDDNOS). Etiologically, of the persons with mental retardation, 18 were diagnosed with fragile X syndrome and 2 with Down syndrome. One of the persons with autism had fragile X syndrome and none had Down syndrome.

Interrater and test–retest reliability and internal consistency of the subscales of the ABI were assessed in a subset of 16 of the parents of persons with autism. All of the persons with autism were male. Their ages ranged from 5.8 to 19.58 years ($M = 11.32$, $SD = 4.57$). Communication age equivalents from the VABS ranged from 0.41 to 8.41 years ($M = 2.36$, $SD = 2.15$). Intelligence test scores (Slosson, 1981) ranged from 2 to 107 ($M = 31.72$, $SD = 22.27$). Of the 16 cases, 3 had total CARS scores that placed them (based on the CARS criteria) in the Mildly–Moderately Autistic

Range (scores of 36, 35, and 31) and the rest fell into the Severely Autistic Range (all scores greater than 36).

Autistic Behavior Interview (ABI) Structure

ABI items are arranged into four broad behavioral areas: Reciprocal Social Interaction, Verbal/Nonverbal Communication Skills, Restricted Interests, and Mood and Arousal Level. There are a total of 28 subscales where each is composed of six items arranged within the subscale according to an ascending hierarchy of ability (for items reflecting appropriate behaviors) or severity (for items reflecting inappropriate behaviors). Scoring for each item in the subscale uses a Likert scale ranging from 0–3: *never* (0), *rarely/emerging* (1), *sometimes/partially* (2), and *often/typically* (3). A manual describing the scoring criteria was provided to the interviewers. The score for each subscale is the total score of the six items. Thus, scores could theoretically range in value from 0 to 18.

Reliability Procedures

For the subset of 16 parents, reliability measures were assessed as follows: One rater interviewed the caregiver/parent and approximately 1 week later, the second rater interviewed the same parent/caregiver to test interobserver reliability. Approximately 1 week after the second interview, the first rater reinterviewed the same informant to assess test–retest reliability. Order of raters was counterbalanced. One interviewer (D.L-J.) was a graduate student in psychology with relatively little experience with persons with autism. The other interviewer (M.K.) was a postdoctoral fellow with training in clinical neuropsychology whose primary experience was with persons with brain injury and/or psychiatric disorders. Interviewers more experienced with autism were excluded in order to provide a reasonable estimate of reliability for persons in the field who might expect to use this instrument to assist in diagnosis. Subscale interrater reliabilities were assessed across raters using the intraclass correlation coefficient and test–retest reliabilities using the Pearson coefficient. Internal consistency was computed for each subscale using the alpha coefficient.

Selection of ABI Items for Neural Network Classification

From the 28 subscales on the ABI, 15 were initially selected to be used in differential classification because of their direct relevance to the

DSM-III-R criteria. These included subscales tapping Reciprocal Social Interaction: (a) reactions to pain, (b) communication of hunger or thirst, (c) empathy, and (d) social interaction; Verbal/Nonverbal Communication Skills and Imagination: (a) eye contact, (b) tactile defensiveness, (c) facial expression, (d) gesture, (e) vocal intonation, (f) imagination, (g) understanding of language, (h) ability to use verbs for communication, (i) verbal perseveration, and (j) topic perseveration; and Restricted Interests: A sum score of the presence, to any degree (Likert score greater than 0), of (a) motor stereotypies, (b) object stereotypies, (c) self-injury, (d) sameness problems, and (e) becoming upset in reaction to changes in routine. This subscale was computed because most of the ABI subscales that tapped each of these areas had low to moderate interrater reliability (see below). This subscale ranged from 0 to 5 and had acceptable interrater reliability (see below).

Subsequent analysis of the intercorrelation matrix for these 15 items revealed that 2, communication of hunger and thirst and ability to use verbs for communication, were highly correlated with several other measures and were therefore dropped from the analysis to avoid redundancy. In addition, empathy and facial expression correlated highly, r(136) = .70, p < .001, as did vocal intonation and understanding of language, r(136) = .81, p < .001. These four measures were therefore combined into two measures, empathy/facial expression and intonation/understanding, by multiplying their respective scores (R. Everly, personal communication). Thus, 11 measures remained for further analysis.

Neural Networks

From the neural network models that were available, a feed-forward network using a back-propagation training algorithm (Rumelhart, Hinton, & Williams, 1986) was selected based on information indicating that this method has a variety of purposes and is very useful for categorization and generalization (Caudill & Butler, 1992; NeuralWare, 1991). The NeuralWorks Professional II Plus program (NeuralWare, 1991) was used for all computations.

Neural networks are computer simulations of interactions among neurons. As shown in Figure 2, a typical feed-forward network consists of at least three layers: (a) an input layer consisting of an investigator-defined number of artificial neurons (or processing elements) analogous to sensory input; (b) one or more "hidden" layers with the first hidden layer receiving connections from the input layer, and (c) an output layer that arrives at a classification based on input from the hidden layer(s). The

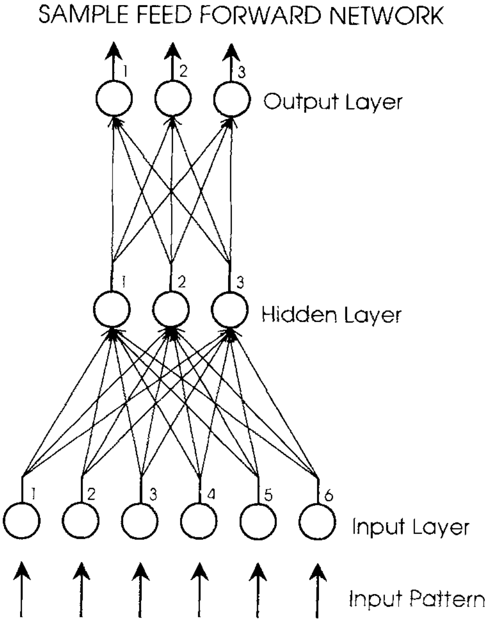

Fig. 2. A typical feed-forward neural network with six input neurons, three hidden neurons, and three output neurons. In this network, all of the input neurons are connected to all of the hidden neurons and all of the hidden neurons are connected to all of the output neurons. These multiple connections allow for distribution of information in the network.

number of neurons in the output layer depends on the number of categories that require classification.

The network "learns" by altering the strength of the connections (connection weights) among interconnected neurons (cf. Hebb, 1949) based on the feedback it receives from the error in its output. These connection weights may be either positive, resulting in excitatory input, or negative, resulting in inhibitory input, to the subsequent neuron(s). Each neuron arrives at a decision to send input to the next neuron based on the summed weighted input from the connections it receives. The magnitude of the output in the hidden and output layers is usually based on a nonlinear, sigmoid function (the transfer function—in this case a hyperbolic tangent function) analogous to the all-or-none response of a neuronal axon. That is, as the weighted input increases, the magnitude of the output from the neuron increases in an S-shaped or ogival manner.

The network is trained by presenting information (e.g., a vector having each of the ABI subscale scores for a given subject) at the input level. Initially, the connection weights among the neurons are random. After input is presented to the network, a signal is generated at the output which is a "guess" of the category to which the subject belongs. This output is compared with the "desired" output determined by an a priori classification defined by the trainer; in this case, autism versus mental retardation. The difference (accumulated over N cases) between the predicted and desired output multiplied by the derivative of the transfer function is defined as the error which is then back-propagated to the hidden and input layers. The connection weights are then altered in an attempt to minimize this error according to a variation of the "generalized delta rule":

$$w'_{ij} = w_{ij} + C_1 * e_i * x_{ij} + C_2 * m_{ij}$$

where w'_{ij} is the updated weight vector for neuron i to neuron j, w_{ij} is the previous weight vector, C_1 is a learning coefficient varying from 0 to 1.0, e_i is the error defined above, x_{ij} is the input to the ith neuron from the jth neuron, C_2 is another learning coefficient and m_{ij} is a momentum term which tends to keep weight changes moving in the same direction despite sudden changes in x_{ij} or e_i (Caudill & Butler, 1992).

This process of changing the connection weights is repeated after every N presentations of randomly presented data until the overall or global error is minimized. Hopefully, the network, at that point, will have "learned" the associations that generate a correct classification. This training process usually requires thousands of case presentations (iterations).

The utility of a trained network is evident only by its ability to *generalize*, that is, to accurately predict the outcome of cases it has never experienced. Generalization is facilitated if the network has many examples of the variety of inputs expected to be associated with a given category and if the network is not too complex for the data. If there are too few data points, the network is too complicated or training is carried out for too many data presentations, the network may learn to categorize by memorizing arbitrary and unimportant features of the training set. Generalization is facilitated, therefore, by adding noise to the layers of the network if the data set is small. This has the effect of slightly changing the input to the connecting neuron each time that same set of data is experienced and therefore decreases the likelihood of memorization of the data set. In addition, periodically testing the network for generalization while training is occurring can help prevent overlearning.

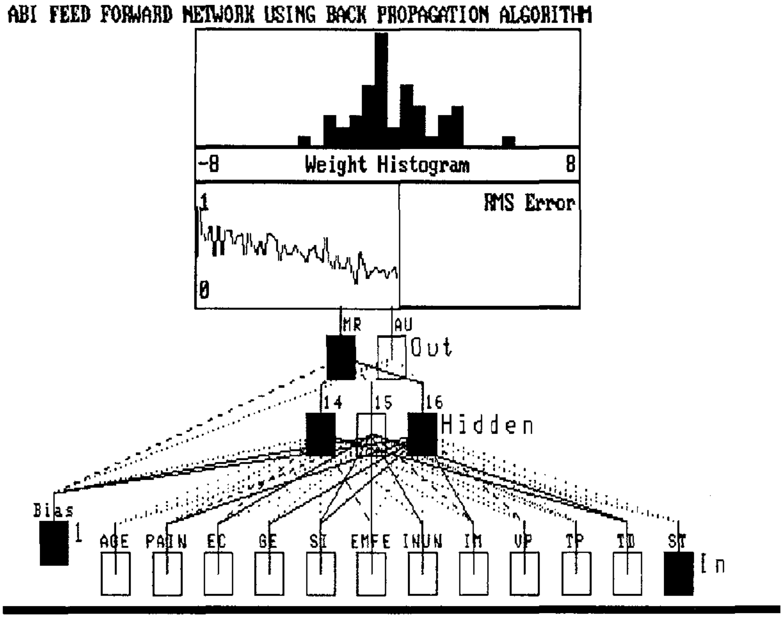

Fig. 3. One of the 13 networks that was successful in both learning and generalizing in the present study. There are 12 inputs identified by name on the top of each input neuron, 3 hidden neurons, and 2 output neurons identified as MR (output for the mental retardation classification) or AU (output for the autism classification). Excitatory connections are indicated by dashed or solid lines and inhibitory connections by dotted lines. See text for discussion of the other features of the network.

Network Model

The model that evolved after a number of different attempts has the following characteristics, as shown in Figure 3:

1. The network has 12 inputs (features)—the 11 measures defined above, and chronological age, since it was felt that many of the subscales will interact with level of development. This provided a subject to variable ratio of over 5, a minimum value typically recommended for multivariate analysis.

2. Three neurons are present in the hidden layer. These hidden layer neurons are thought to form "internal representations" of the input data which are "emergent characteristics" (Rumelhart, Hinton, & Williams, 1986) that arise after extended experience with the associations to be learned. In other words, these hidden neurons do not form logical decision rules but, instead, function as feature pattern detectors. Three hidden neurons were found to provide good generalization with a relatively small number of large connection weights in the network, a desirable outcome.

3. Two neurons are present in the output layer with one serving as the mental retardation classifier and the other as the autism classifier. The decision to classify in one direction or the other is, itself, a function of the joint weighted input received from all of the hidden layer neurons.

4. A "bias" processing neuron is present and serves as a means of adjusting the location(s) of the complex multidimensional space functions that separate groups from each other. It is functionally analogous to the constant term in the classification functions provided in linear discriminant analysis (D. J. Shazeer, personal communication, April 25, 1992).

Training and Testing the Network

Since we did not have a large data set available for cross-validation, a "jackknife" or "leave k out" procedure was used to train and test the network (Sklansky & Wassel, 1981), a procedure also used by Weinstein et al. (1992) in their study of cancer drugs. Of the total of 138 cases, 5 cases in each group (a total of 10 cases) were set aside for generalization testing leaving 64 cases per group for training purposes. These 5 were generated by sorting the groups on age and selecting every 13th case in each group. This assured that both the training and test sets would consist of an equivalent range of ages.

Initially, the connection weights among the neurons were randomized to range from −0.1 to +0.1 using a random number seed of 57. Training was carried out by presenting a vector having each of the scores for a given subject at the input level. This produced a linear output based on the relative magnitude of each of the input neurons (scaled to range in value from −1.0 to 1.0). The output of each of the input neurons was sent to each of the neurons in the hidden layer weighted by the size of the connection weights between the input and hidden layer neurons. The hidden layer neurons then produced an output based on the size of the input. The form of the function relating output to input *at this level* was a hyperbolic tangent, S-shaped function. Output from the hidden neurons was then sent to the output neurons where a similar nonlinear transformation took place. This defined one iteration. This output was compared with the desired output, −1,1 or 1,−1 corresponding to autism or mental retardation, respectively. As defined above, in an attempt to minimize the overall error, the connection weights were altered, by back-propagating the error to the hidden layer neurons and then to the input layer neurons. The connections among the layers were altered based on that error. This process of changing the connection weights was repeated after every 32 presentations of data randomly selected from the overall data set.

Since some networks have a tendency to overlearn the data set if training is carried out for too long a time, as noted above, the following procedure was followed: The network was trained for 5,000 iterations or case presentations to allow for initial learning of the data set and then stopped. At this point, training was then restarted and, every 150 iterations, generalization to the test set was evaluated. Those points in training where generalization evidenced improvement (in terms of a reduced root mean square error) over previous tests were saved and the best result used in later calculations of generalization. This procedure continued for 30,000 iterations or until generalization failed to improve over 180 successive tests (about 20 min on a 20 megahertz 386-SX computer with math coprocessor).

This test set was then put back into the training set and a new test set was set aside by taking every 13th case starting at the previously identified cases plus one in each group. *The network was then initialized* as described above and trained on this new data set and tested on the new test set. This process was repeated 11 more times until a total of 130 cases had been tested for generalization.

Because learning in a network is a computational and not a statistical process (in the usual sense of the term), it is likely that repetition of the above procedure with the same initial parameters will lead to slightly different results. Therefore, in order to estimate the overall performance of this network anatomy, the procedure described above was repeated and the average error in the training and test sets computed.

During training, the learning rate (learning coefficient C_1) for the hidden layer was initially set to a level of 0.30 for the first 10,000 iterations in order to rapidly approach a solution and then slowed to 0.15 after this in order to narrow in on the appropriate solution. The learning rates on the output layer were one half of the middle layer rates. In addition, to facilitate generalization, uniform random noise (±40%) was added to the hidden layer neurons for the first 10,000 iterations and then decreased to 20% for the rest of the training period. The presence of noise and its decrease across blocks of iterations is also recommended for small data sets to avoid "local minima" in the error that is to be minimized.

Discriminant Function Analysis

For comparison, the same 13 sets of data were presented to a simultaneous as well as a stepwise (F to enter=1.0) linear discriminant function analysis (Statsoft, 1991) for training and testing purposes. Independent variables consisted of the same 12 variables indicated above with classification of autism or mental retardation as the dependent variable.

RESULTS

Reliability Assessment

As shown in Table I, test–retest reliabilities were usually higher than interrater reliabilities. Correlations that were not significant at the .05 level (one-tail) are indicated (see Footnote *a*, Table I). The modal interrater reliability range peaked at .60–.69. The modal test–retest reliability range was .80–.89. Depressive behavior, which required that signs of depression be present for a minimum of 10 days, was rarely reported in this sample. At least two subscale reliabilities including eye contact and negative reactions to sounds appeared to be relatively low because of range restriction, that is, there was relatively little variation in these measures across subjects as shown in Table I by the semi-interquartile ranges for raters 1 (R1) and 2 (R2). Five additional subscales had uniformly poor reliability despite adequate range: seeking approval, aversion to social interaction, vocal and object stereotypies, and hyperactivity. Of these five, seeking approval continued to show poor reliability even on retest. Raters had problems with the severity dimension used to assess the persistence of stereotypies, that is, asking how soon the behavior starts after it has been stopped. These measures needed to be redefined. The composite measure of stereotypy defined above: a sum score of the presence, to any degree (Likert score greater than 0), of (a) motor stereotypies, (b) object stereotypies, (c) self-injury, (d) sameness problems, and (e) becoming upset in reaction to changes in routine, had acceptable interrater reliability (intraclass $r = .80$). Preliminary multidimensional scaling of the data suggested these Restricted Interest subscales do, in fact, cluster together.

Internal consistency measures were uniformly high (>.80). The only exceptions were communicating hunger and thirst (.73), social interaction (.77), imagination (.58), and facial expression (.39). The subscale of facial expression tapped many different aspects including smiling, interest, fear, surprise, and so forth, and the poor alpha coefficient supports other research indicating that descriptors of emotions are not unidimensional (e.g., the bipolar dimensions of pleasure–displeasure vs. high arousal–low arousal described by Russell, 1989).

Neural Network and Discriminant Analysis Results: Training

Figure 3 shows a typical trained network along with several graphs. In the upper center is a histogram of the connection weights in the network and underneath it is a display of the root mean square error output over blocks of iterations with an abscissa ranging from 0 to 32,000. The latter

Table I. Interrater (IR) Reliability, Semi-Interquartile Ranges for Raters 1 (R1) and 2 (R2), Test–Retest (TRT) Reliability, and Alpha Coefficients

	IR	R1	R2	TRT	α
Reciprocal social interaction					
Pain	.75	6	7	.87	.88
Hunger	.60	6	6	.62	.73
Eye contact	.25[a]	3	3	.73	.81
Empathy	.73	5	6	.75	.84
Social interaction	.53	3	7	.82	.77
Seeking approval	.00[a]	9	7	.35[a]	.89
Agonistic behavior	.64	6	2	.72	.89
Verbal/nonverbal communication					
Tactile defensiveness	.69	6	6	.71	.98
Aversion to social interaction	.26[a]	10	6	.65	.89
Facial expression	.61	4	2	.68	.39
Gesture	.59	2	3	.78	.68
Intonation	.88	9	6	.84	.81
Imagination	.56	4	3	.70	.58
Understanding	.68	9	12	.86	.93
Verbal perseveration	.64	11	8	.96	.96
Generalization of verbs	.91	10	7	.90	.90
Topic perseveration	.69	2	0	.83	.95
Restricted interests					
Motor stereotypies	.49	8	11	.81	.99
Visual stereotypies	.58	10	2	.59	.99
Vocal stereotypies	.28[a]	10	6	.48	.98
Object stereotypies	.13[a]	10	10	.45	.97
Sameness	.80	9	4	.89	.98
Disturbance in routines	.64	10	8	.79	.98
Self-injury	.58	4	2	.76	.89
Mood and arousal level					
Depressive behavior	—	4	—	—	—
Agitation	.45	10	8	.52	.99
Reactions to sounds	.30[a]	4	0	.48	.91
Hyperactivity	.27[a]	9	12	.82	.98

Table II. Training Results of the Neural Networks and the Discriminant Analyses for Persons with Autism (AUT) or Mental Retardation Without Autism (MR)

	Neural network			Discriminant analysis	
	First pass	Replication	*M*	Simultaneous	Stepwise
% correct AUT	95	95	95	89	89
% correct MR	87	84	86	81	80
No. of iterations	12,292	8,484	10,388		

"learning curve" decreases with training, with periodic increases in error as mistakes are made as a result of gradual changes in the connection weights. Eventually, the error decreased to a level of about 0.20 where it stopped because it was at this point that maximum generalization to the test set occurred.

Average percentage accuracy across the 13 data sets for the two network repetitions are shown in Table II and compared with the discriminant analyses. As shown, the networks classified 95% of the cases with autism and 86% of the cases with mental retardation during training after a joint average of 10,388 iterations. By contrast, the 13 simultaneous linear discriminant analyses classified a total of 89% of the cases with autism and 81% of the controls who were mentally retarded (for the entire data set presented at once, Wilks's lambda was .51, $F(12, 126) = 10.10$; $p < .0001$). The 13 stepwise linear discriminant analyses produced identical results and classified a total of 89% of the cases with autism and 80% of the controls who were mentally retarded (for the entire data set presented at once, Wilks's lambda was .51, $F(8, 130) = 15.09$; $p < .0001$).

Thus, as we would predict, there was a 6% improvement in the training set data for the autistic sample which was significantly different from the percentage identified by both the simultaneous and stepwise discriminant analyses (Wilcoxon $T = 13$, $p = .012$, one-tail test for each comparison). Similarly, there was a 5 to 6% improvement in the training set data for the mentally retarded sample which was significantly different from the percentage identified by simultaneous (Wilcoxon $T = 17$, $p = .023$, one-tail test) and stepwise discriminant analysis (Wilcoxon $T = 8$, $p = .008$, one-tail test).

Generalization Testing

Results of generalization testing are shown in Table III. The neural networks correctly predicted, on average, 92% (±3%; 95% confidence

Table III. Generalization Results of the Neural Networks and the Discriminant Analyses for Persons with Autism (AUT) or Mental Retardation Without Autism (MR)

	Neural network			Discriminant analysis	
	First pass	Replication	*M*	Simultaneous	Stepwise
% correct AUT	98	95	97	85	83
% correct MR	89	85	86	78	78

interval) *of cases they had never seen*—97% of the persons with autism (±4%, 95% confidence interval) and 86% (±7%; 95% confidence interval) of the persons with mental retardation, virtually the same as the training set data indicating no shrinkage in predictability. The simultaneous discriminant analyses fared less well predicting 82% (±6%; 95% confidence interval) of the overall cases—85% (±7%; 95% confidence interval) of the persons with autism and 78% (±9%, 95% confidence interval) of the persons with mental retardation. The stepwise analyses led to the same results and generalized to 83% of the persons with autism and 78% of the persons with mental retardation. Thus, there was a 4% shrinkage in predictability for the discriminant analyses which was not evident in the networks. Fletcher, Rice, and Ray (1978), using a random number set, have found that a 5:1 ratio of subjects to variables (as in the present study) produces a shrinkage in predictability in cross-validation of 9 to 12% in samples of 50 to 250 per group. The shrinkage factor here was lower.

Since the nonautistic group had a much lower correct classification and prediction rate than the autistic group, the predicted misclassified cases were examined. In the seven cases that both network passes agreed on, six had a diagnosis of PDDNOS. Either our diagnosis was incorrect in these cases or the network did not have the necessary information to make this distinction.

As predicted, the overall classification accuracy for the network models was significantly superior for the groups with autism (McNemar $\chi^2 = 7.11$, $p = .004$, one-tail test) and mental retardation (McNemar $\chi^2 = 3.13$, $p = .039$, one-tail test).

Interpretation of Hidden Layer Connection Weights

Interpretation of the connection weights of hidden layer neurons is not a straightforward process, particularly if the number of hidden neurons is large. In such cases, it is helpful to probe the network with selected stimuli

Table IV. Connection Weights from Input to Hidden Neurons

Input	Hidden neuron 14	Hidden neuron 15	Hidden neuron 16
Bias	–0.78	1.15	1.66
Age	–1.81	–3.27	–0.57
Pain	–0.96	0.88	1.50
Eye contact	0.55	5.23	–1.76
Gesture	–2.50	–0.17	2.87
Social interaction	–0.33	1.24	2.26
Empathy/face expression	0.16	3.04	–0.16
Intonation/understanding	–0.22	2.35	–1.48
Imagination	–0.08	0.78	2.93
Verbal perseveration	–0.20	–0.15	0.80
Topic perseveration	–0.46	–0.03	–2.22
Tactile defensiveness	2.54	2.94	–0.67
Stereotypy	–0.41	–2.25	–1.45

and see which of the neurons are activated (Caudill & Butler, 1992). In the present case, fortunately, the number of hidden neurons was small and interpretation was rendered somewhat easier.

Table IV shows the connection weights of the inputs entering the three hidden neurons of the network for Data set 7. This network was saved after 16,700 iterations had been completed (other nets produced similar results). In this network, the weighted connections from each of the hidden layer neurons to the output designating "autism" were negative (not shown in the table). Therefore, the autism category would be excited at the output layer if the summed weighted input to that neuron from the hidden layer were, itself, negative (since a negative input from the hidden layer neurons multiplied by negative connection weights equals a positive activation). Therefore, to interpret the hidden layer connection weights insofar as they indicate prototype behavior patterns that are important for a categorization of autism, it is important to highlight those inputs that have negative connection weights and to minimize the importance of those inputs that have positive connection weights.

Almost all of the inputs to Hidden Neuron 14 were negatively weighted except for eye contact, empathy/facial expression, and tactile defensiveness and the weights for the first two of these inputs were relatively weak. Similarly, the connection weights for almost all of the other negatively weighted items were also weak except for age, pain, and gesture. Therefore, this neuron would yield a negative output if the inputs that were negatively weighted exceeded the inputs that were positively weighted. This would occur with a behavior pattern characterized by increased age, some verbal and imaginative skills (communication of pain, gesture, social interaction, intonation/understanding, and imagination), weak nonverbal social skills (eye contact, empathy/facial expression), minimal tactile defensiveness, and weakly emphasizing some perseverative language (verbal and topic perseveration) and stereotyped behavior. This behavior pattern is reminiscent of Wing and Attwood's (1987) "Passive Group."

Almost all of the inputs to Hidden Neuron 15 were positively (and strongly) weighted except for age, gesture, verbal and topic perseveration, and stereotypy with very weak connection weights for gesture, verbal, and topic perseveration. Therefore, this neuron would yield a negative output with a behavior pattern characterized by increased age, very limited verbal and imaginative skills (communication of pain, social interaction, intonation/understanding, and imagination), very weak nonverbal social skills (eye contact, empathy/facial expression), minimal tactile defensiveness but with stereotyped behavior, and, to some extent, perseverative verbalization (verbal perseveration). This behavior pattern is reminiscent of Wing and Attwood's (1987) "Aloof Group."

About one half of the inputs to Hidden Neuron 16 were negatively weighted except for communication of pain, gesture, social interaction, imagination, and verbal perseveration. Therefore, this neuron would yield a negative output with a behavior pattern characterized by increased age, the presence of some social and verbal skills (eye contact, intonation/understanding, and to a weak extent, empathy/facial expression), a reduced degree of other nonverbal and imaginative skills (communication of pain, gesture, social interaction, and imagination), but with sophisticated perseverative language (topic perseveration), tactile defensiveness, and stereotyped behavior. This behavior pattern is reminiscent of Wing and Attwood's (1987) "Active-but-Odd Group."

All of these prototypes have in common a core pattern of poor receptive and expressive nonverbal and imaginative skills along with perseverative behavior (verbal or nonverbal).

DISCUSSION

These results are encouraging and clearly require replication. They suggest that when provided with relevant information such as that generated by the ABI, neural network models can be more powerful than more traditional models in their ability to categorize and generalize a behaviorally defined syndrome. The fact that shrinkage in predictability was not a problem for the networks suggests that they were more tolerant of variation in the input data than the discriminant analyses. It should be noted, however, that our results are based on a relatively small sample and our observations on the emergent characteristics of the hidden neurons are likely to be unstable and, perhaps, sample- and rating-scale-dependent. For a net to learn the range of variability to be expected in caregiver reports of persons with autism versus persons with retardation, at a given age, sex, or etiological diagnosis, requires substantially more cases. It remains to be seen if other investigators can achieve similar classification results using other subjects and/or assessment materials.

We point out that the particular network anatomy and algorithm chosen in the present study is not the only type available. Rather than arriving at a category by plurality agreement among hidden neurons, some networks allow these neurons to compete with one another (winner take all) in terms of their success in making accurate classifications. These include Learning Vector Quantization networks (Caudill & Butler, 1992) and Grossberg's Adaptive Resonance models (cf. Carpenter & Grossberg, 1991) and may also be useful in identifying unique subgroups under the umbrella term of autism.

If our results can be replicated with the same or other anatomies, neural network methodology may help to resolve some important diagnostic issues:

1. Do different investigators really emphasize different features in classifying persons as having autism? Comparison of the ability of a variety of networks to classify and generalize classification of persons as autistic based on clinical judgments of a variety of different experienced investigators along with the emergent characteristics of their hidden neurons may help to determine if these theoretical impressions are more apparent than real.

2. Are there observable core characteristics of autism that can be reliably identified? Even though the networks were not presented with prototypical subgroups of autism, it was of some interest that the emergent characteristics of the hidden neurons in this sample bore some similarity to Wing and Attwood's (1987) subtypes. Analysis of the emergent characteristics

of neural networks based on a large sample of persons with autism using a variety of experts to assist in diagnosis may help to elucidate this issue. The ability of such networks to generalize to new samples will test the stability of these core characteristics.

3. Can inexperienced investigators and clinicians accurately diagnose autism using currently available assessment tools? If a subset of persons can be agreed upon by a number of experts as clearly autistic and other relevant and agreed upon diagnostic groups can be identified (e.g., mental retardation not associated with autism, PDDNOS), then a network could, theoretically, be trained to make this discrimination given an appropriate assessment tool(s), such as the ABI. The trained and tested network could then be saved as a simple program. Any clinician or investigator could use the recommended assessment tool(s) to input data to the network and the network program could then provide a probability of assignment of that person to the relevant diagnostic category. Further, having identified unique patterns of behavior associated with a variety of disorders may also help to classify those borderline cases that do not appear to fit into one pattern or the other.

4. Do persons with autism show different and predictable behavior patterns depending on their developmental level, etiology, sex, drug state, or concurrent psychiatric or neurological disorder? Assuming enough subjects could be obtained for training and testing, this question could be answered with a theoretically high level of reliability. If valid patterns are evident, behavioral features may therefore be readily predictive of other important factors.

We have argued here that the diagnostic process is more akin to pattern recognition than it is to adding together a group of symptoms as in a rule-based expert system. As such, it may be necessary to use systems that mimic this process, such as neural networks, in order to provide stable and comprehensive descriptions of behavioral types and subtypes for both clinicians and researchers. In short, we feel that neural network modeling of the diagnostic process has validity for identifying and understanding persons not only with autism but with *any* condition in which a unique behavioral pattern is expected (e.g., learning disability, schizophrenia, tic disorders). The generality of this statement is, of course, an empirical issue.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern recognition by self-organizing neural networks.* Cambridge, MA: MIT Press.

Caudill, M., & Butler, C. (1992). *Understanding neural networks.* (Vols. 1 & 2). Cambridge, MA: MIT Press.

Creak, M. (1961). Schizophrenia syndrome in childhood: Progress report of a working party. *Cerebral Palsy Bulletin, 3,* 501-504.

Fletcher, J. L., Rice, W. J., & Ray, R. M. (1978). Linear discriminant function analysis in neuropsychological research: Some uses and abuses. *Cortex,* 14, 564-577.

Freeman, B. J., Ritvo, E. R., Yokota, A., & Ritvo, A. (1986). A scale for rating symptoms of patients with the syndrome of autism in real life settings. *Journal of the American Academy of Child Psychiatry, 25,* 130-136.

Goodman, P., Kaburlasos, V., Egbert, D., Carpenter, G., Grossberg, S., Reynolds, J., Hammermeister, K., Marshall, G., and Grover, F. (1992). *Fuzzy Artmap neural network prediction of heart surgery mortality.* Poster presented at research conference entitled Neural Networks for Learning, Recognition and Control, Wang Institute of Boston University, May 14-16.

Grossberg, S. (1984). Some normal and abnormal behavioral syndromes due to transmitter gating of opponent processes. *Biological Psychiatry,* 19, 1075-1118.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child 2,* 217-250.

Krug, D. A., Arick, J., & Almond, P. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry, 21,* 221-229.

Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., & McLennnan, J. (1989). Autism diagnostic interview: A standardized investigator-based instrument. *Journal of Autism and Developmental Disorders, 19,* 363-385.

NeuralWare. (1991). *Neural computing.* Pittsburgh, PA: Author.

Ornitz, E. M., & Ritvo, E. R. (1968). Perceptual inconstancy in early infantile autism. *Archives of General Psychiatry, 18,* 76-98.

Parks, S. L. (1983). The assessment of autistic children: A selected review of available instruments. *Journal of Autism and Developmental Disorders, 13,* 255-267.

Rapin, I. (1987). Searching for the cause of autism: A neurological perspective. In D. J. Cohen & A. M. Donnellan (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 710-717). Silver Spring, MD: Winston.

Reiss, S., Levitan, G., & Szyszko, J. (1982). Emotional disturbance and mental retardation: Diagnostic overshadowing. *American Journal of Mental Deficiency, 86,* 567-574.

Ritvo, E., & Freeman, B. J. (1978). National society for autistic children definition of the syndrome of autism. *Journal of Autism and Childhood Schizophrenia, 8,* 162-167.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 45-76). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Russell, J. A. (1989). Measures of emotion. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, research and experiencek* (Vol. 4, pp. 83-111). San Diego: Academic Press.

Rutter, M. (1978). Diagnosis and definition of childhood autism. *Journal of Autism and Childhood Schizophrenia, 8,* 139-161.

Rutter, M., Bartak, L., & Newman, S. (1971). Autism: A central disorder of cognition and language. In M. Rutter (Ed.), *Infantile autism: Concepts, characteristics and treatment* (pp. 148-171). London: Churchill-Livingstone.

Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The childhood autism rating scale.* Los Angeles: Western Psychological Services.

Sevin, J. A., Matson, J. L., Coe, D. A., Fee, V. E., & Sevin, B. M. (1991). A comparison and evaluation of three commonly used autism scales. *Journal of Autism and Developmental Disorders, 21*, 4, 417-432.

Sklansky, J., & Wassel, G. N. (1981). *Pattern classifiers and trainable machines.* New York: Springer-Verlag.

Slosson, R. L. (1981). *Slosson intelligence test.* Los Angeles: Western Psychological Services.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales. Interview edition. Survey form manual.* Circle Pines, MN: American Guidance Service.

Statsoft (1991). *CSS: Statistica.* Tulsa, OK: Author.

Volkmar, F. R., Bregman, J., Cohen, D. J., & Cicchetti, D. (1988). DSM-III and DSM-III-R diagnoses of autism. *American Journal of Psychiatry, 145,* 1404-1408.

Wadden, N. P. K., Bryson, S. E., & Rodger, R. S. (1991). A closer look at the autism behavior checklist: Discriminant validity and factor structure. *Journal of Autism and Developmental Disorders, 21*, 529-541.

Weinstein, J. N., Kohn, K. W., Grever, M. R., Viswanadhan, V. N., Rubinstein, L. V., Monks, A. P., Scudiero, D. A., Welch, L., Koutsoukos, A. D., Chiausa, A. J., & Paull, K. D. (1992). Neural computing in cancer drug development: Predicting mechanism of action. *Science*, 258, 447-451.

Wing, L., & Attwood, A. (1987). Syndromes of autism and atypical development. In D. J. Cohen, & A. M. Donnellan (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 3-19). Silver Spring, MD: Winston.