©Copyright 2005
Anne Thissen-Roe

Adaptive selection of personality items to inform a neural network predicting job performance

Anne Thissen-Roe

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2005

Program Authorized to Offer Degree:
Psychology

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Anne Thissen-Roe

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

_____

Earl Hunt

Reading Committee:

_____

Earl Hunt

_____

David Corina

_____

John Miyamoto

Date:_____

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106 -1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

Adaptive selection of personality items to inform a neural network predicting job performance.

Anne Thissen-Roe

Chair of the Supervisory Committee:
Professor Emeritus Earl Hunt
Department of Psychology

Connectionist or "neural" networks, developed as a model of cognition, are also a general statistical model with practical applications. Adaptive testing, traditionally based on item response theory, is a way to improve the efficiency of a test. A hybrid system is developed that captures the main advantages of both technologies: the modeling flexibility of a neural network, and the efficiency gains of adaptive testing. A prototype is implemented for the case of a personality assessment used to predict job tenure at a national retail chain. Applicants' assessment and subsequent employment data are used to demonstrate the prototype's effectiveness.

TABLE OF CONTENTS

Page
List of Figures....ii
List of Tables....iii
1. Introduction....1
1.1. Employment Testing....3
1.2. Neural Networks....12
1.2.1 Network Architectures....18
1.2.2 Useful Properties....24
1.3. Computerized Adaptive Testing....28
1.4. Subtests and Scoring Considerations....36
2. A hybrid selection algorithm....42
2.1. Example 1: All data present but one item....45
2.2. Example 2: Two items missing....47
2.3. Example 3: Many items missing....48
2.4. Error of measurement, and a candidate selection algorithm....50
2.5. Uncertainty in latent trait values....52
2.6. A better item selection algorithm....56
2.7. Modularity....58
3. Implementation and program structure....60
3.1. Applicant Interface....64
3.2. Sequencer....65
3.3. Logs....66
3.4. Item Selection Rule....67
3.5. Scoring Rule....68
3.6. Latent Trait Structure....69
3.7. Neural Network....71
3.8. Optimization....72
4. A simulation study....75
4.1. Source tests....76
4.2. Neural Network....79
4.3. Method....81
4.4. Results....81
5. Conclusion....85
References....87

LIST OF FIGURES

Figure Number Page
1. One node of a perceptron.....22
2. Architecture of the program, from a process standpoint.....61
3. Architecture of the program, from a data flow perspective.....63
4. A screen phone.....77

LIST OF TABLES

Table Number Page
1. Mean absolute difference from the "all data" condition.....83
2. Mean standard error of measurement as reported by the test.....83
3. Correlation with "all data" condition.....83

ACKNOWLEDGEMENTS

The author wishes to thank Unicru, Inc. for providing the data and other resources necessary to complete this project. The author also wishes to express her gratitude to the numerous individuals at the company who went out of their way to be helpful. The Science Team were forthcoming with references and shared expertise in the early stages of this research; the Modeling Team put in extra effort to provide data within the time constraints of the project. Your behaviors do you credit.

Dave Thissen made a number of insightful comments during the development stages of this project, which were and are still deeply appreciated.

Thanks are also due to several individuals at the University of Washington. David Corina, Jim Minstrell, John Miyamoto and Bryan D. Jones were patient, accommodating, thorough and supportive committee members. Sandi Dormont and Nancy Kenney made several "diving catches" to ensure that deadlines were met and procedures were followed, despite two hundred miles' distance. Finally, the author's success is in significant part attributable to the guidance of her advisor and committee chair, Earl "Buz" Hunt.

1. Introduction

Consider the problem of hiring a new employee when several candidates are available. It is preferable to get the best available candidate, or at least, to avoid the worst. The time and effort spent evaluating candidates have real costs to a business, but hiring the wrong person may lead to firing that person and starting the process over. The wrong candidate may also steal from the business, be unsafe and risk injury for which the business is liable, or expose the business to costly lawsuits.

A brief assessment related to the job is a way of selecting an above-average candidate more than half of the time. Assessments have been used in employee selection for decades, but computers can make them even more efficient. With the automation of the job application, data entry is removed from the process. At the same time as it records applicant data, the computer can score the assessment, and evaluate the candidate according to strict rules. Network transmission permits centralized storage and continuous or routine monitoring of applications submitted at many locations. This process has a number of beneficial side effects, from reduction of paperwork to reduction of discrimination.

Any valid assessment can improve the quality of the hiring decision over none, including procedures such as interviews that we may not think of as assessments, but also more formal tests. Technological sophistication may improve the quality of the assessment, an improvement which is passed along to the hiring decision. Different technologies address different problems, but may be difficult to use in conjunction with

each other. This paper discusses a case of joining two technologies with different functions. Specifically, a neural network is a general statistical model of the predictive relationship between assessment and outcome, which allows for nonlinear interactions between measures within a broad assessment. Adaptive item section can make a test more efficient while minimizing loss of information. The goals of the two methods are not incompatible, but using adaptive selection with neural net prediction requires careful thought.

This paper develops a novel method of adaptively selecting items to be used as inputs for a predictive neural net. The available data are assumed to be multidimensional, nonlinearly interacting, and variable in utility. It is further assumed that there is a real cost in time and money associated with gathering each piece of information. The method is modular; any of several components can be replaced with a different mathematical technique provided certain constraints are met. Unlike earlier systems of computerized adaptive testing, in which scoring and item selection were strongly integrated, this method can be easily adapted to alternative measurement models. In effect, by combining adaptive testing methods with neural networks, I lay the basis for a methodology of testing that is more flexible, powerful and efficient than current techniques.

The remainder of Chapter 1 will briefly review the background information and prior research which contribute to this hybrid method in four areas: employment testing in general, neural networks, computerized adaptive testing, and assessment structure determination for the type of assessment to be used. Chapter 2 introduces the new method in mathematical terms; Chapter 3 explains how it may be implemented in a computer

program. In Chapter 4, the method is implemented for the case of an assessment used to predict job performance at a national retail chain. Actual applicant and employment data are used to demonstrate the method's function.

1.1. Employment Testing

Employees differ. This is assumed to be self-evident. There are qualities of the employee, as well as of the work and the work environment, that lead to different outcomes after hire, such as productivity, positive behaviors, off-task behaviors, workplace theft and even violence. The overarching purpose of all predictive methods is to anticipate one or more of these outcomes in an applicant before hiring, so that a negative outcome may be avoided or a positive outcome achieved.

Various attempts to predict employee behaviors have focused primarily on predicting two components: competence to do the job, and inclination to do the job. Performance measures may be separated into measures of maximal performance, under which the employee is particularly motivated for the testing period, and typical performance, which reflects both ability and inclination under ordinary conditions (see, e.g., Ployhart et al, 2001; Turner, 1978). Which type of performance is important may depend on particular job conditions. For example, a cash register operator can be slow most of the time and still be considered a good employee, if he picks up the pace to keep up with busy times. Estimating both types of performance, however, calls for knowledge

of both the employee's ability and personality. An assessment may predict one or the other, or both.

The most common class of pre-employment assessment is not what an applicant might think of as a test at all. A fair amount of biographical information is gathered about a job applicant for administrative purposes, and this "biodata" may be used opportunistically to predict success or misbehavior on the job. Biodata may include identifying information, demographic information, information about the applicant's employment history, information about education or credentials, or information about conditions such as veteran status. This is a broader definition than used by Schmidt and Hunter (1998), who distinguish inquiries into job experience and education from the set of all other questions which may be asked. This distinction may not be salient to an applicant when biographical measures used are widely varied.

Biodata may be used to screen applicants quickly for minimum qualifications, such as possession of necessary documents or being old enough to legally work. It may be disregarded for legal or ethical reasons, such as to avoid unfair discrimination against groups, but retained in order to track company demographics, to receive tax credits, or simply to pay the employee. Finally, biodata may be useful in assessing an applicant's competence to do a job, through credentials or job history, and an applicant's behavioral tendencies, also through employment history. Having held a series of related jobs may be a good sign, but getting fired from each one is probably not.

In a meta-analysis across numerous samples and several specific criterion measures, Schmidt & Hunter (1998) give a validity of 0.35 for biodata in predicting job

performance, and lower validities for job experience, educational level, and a measure of training and experience. It is difficult to accept such a value without further qualification, as the utility of biodata no doubt reflects the choice of biodata. Biodata may act as surrogates for constructs such as general mental ability or ambition, which may be measured more specifically.

For the purposes of this paper, it is assumed that some biodata will be collected during the process of application, in order to be passed on to the hiring manager or payroll office, and it may or may not be opportunistically used. In the interests of maintaining applicant dignity and privacy, additional life history inquiries, such as the parental discipline items referenced by Schmidt & Hunter (1998), will not be considered.

A more traditional type of pre-employment assessment is the skills test, and it has close cousins in the knowledge test and the work sample. This group of tests involves direct measurement of the applicant's preparation to do the job. A work sample, for instance, is a rated performance of a selection of job tasks. While the applicant may be more motivated than the hired employee, a demonstration of skill or knowledge still predicts best performance. Predictive validities of 0.54 for work samples and 0.48 for job-related knowledge tests were found in meta-analysis, much higher than the validity of 0.18 given for number of years of experience alone (Schmidt & Hunter, 1998).

Skills tests and work samples are not applicable to untrained or inexperienced workers, nor are they good for "unskilled" jobs, where most of the population possesses the necessary skills or can easily learn them. They are most appropriate to skilled crafts such as carpentry, butchery, welding, and mechanical repair. Similarly, knowledge tests

are only applicable when the applicant has had training, education or experience which is pertinent to the job and not near-universal. (Schmidt & Hunter, 1998)

A second class of test is the ability or aptitude test. These tests can be used with applicants who are expected to be trained in job-specific skills after they are hired. While there are many possible ability tests, including ones to measure physical characteristics such as visual acuity or strength, the most common ability tests measure either general or specific mental abilities.

General mental ability tests have been shown to predict how fast and how well an employee learns a job (Hunt, 1995). Schmidt and Hunter (1998) found validities ranging from 0.23 to 0.58 depending on the complexity of the job, leading the authors to conclude that tests of general mental ability were the most valid and least costly of all broadly applicable selection procedures. The more complex the job, the higher the validity. Over the long term, general mental ability was more important than years of experience, and correlated with skills tests and work samples (Schmidt & Hunter, 1998; Hunt, 1995).

Tests of specific mental abilities, such as spatial ability, memory, and reasoning, are also used in practice. These tests typically load heavily on a general ability factor, but contribute some unique variance (Carroll, 1993).

In low-complexity jobs, where competence to do the job can generally be assumed, the relative value of inclination to do the job increases. Motivation may come from both internal and external influences. Some influences are stable, including expectations of consequences, perceived norms, interests, and personality traits. Others are affected by day to day conditions and may be difficult to predict.

The measurement of personality traits in a work context has been extensively discussed. The set of personality traits that are relevant to job performance is distinct from the set of traits which together fully describe a person. Although many researchers are familiar with small sets of broad personality traits which characterize individual differences in a general sense, such as the Big Five, these factors are sometimes considered to be the top level of a hierarchical model. A broad factor such as Conscientiousness, when closely studied, encompasses related but distinguishable components such as achievement orientation and diligence. More than one level of that hierarchy is of use in the context of employment testing.

Tests of conscientiousness, in its Big Five form, have proven useful for selecting employees. Conscientiousness has a direct, rather than moderated, relationship with job performance, and may predict integrity, responsibility, honesty and reliability, all components of inclination to do a job (Matthews & Deary, 1998; Schmidt & Hunter, 1998; Barrick & Mount, 1991). Specific integrity tests have been used to reduce the likelihood of counterproductive behavior on the job, and may have a higher correlation with performance than broad conscientiousness tests (Schmidt & Hunter, 1998). Not all integrity tests are equal. They may be overt or covert, the latter being closer to tests of the conscientiousness trait (Wanek et al, 1998).

Some personality attributes are useful for selecting employees for particular classes of jobs, but not all jobs. Managers and salespeople both have jobs that call for interaction with new people on a regular basis, an aspect of the job which is either not present or not prominent in many other professions. For these professions, extraversion is

predictive (Barrick & Mount, 1991). Extraversion has components of sociability and ambition, but also tends to reflect general activity level, any of which might be expected to influence performance on some jobs. Scarborough (2005) reports effects of several extraversion-related constructs, including assertiveness and the expectation that one can influence others, on the performance of employees making sales calls. The same study, however, found an effect of emotional resilience, which contradicts earlier findings of no effect of emotional stability (Barrick & Mount, 1991).

It may be inferred from the above discussions that "job performance" is not a trait or behavior, but rather a composite of behaviors influenced by a potpourri of traits. While ability measures may have positive manifold, personality measures are not necessarily correlated with each other or with ability. The predictions to be made by the system described in this paper are further complicated. Job tenure is not, strictly speaking, a performance measure. Tenure may be defined by performance, in that unsatisfactory performers may be fired, but it may also be limited by the employee's comfort with the work and environment. Comfort may or may not be related to performance. There are also more general issues concerning criterion measures, which set the stage for the use of sophisticated statistical models such as neural networks.

As a measure validates or fails to validate against a criterion, so does the theory by which it was developed or chosen. Because of the time scale and stakes involved, experimental manipulations are limited; laboratory conditions can generally not adequately approximate a long-term job environment. Although some manipulations are possible (such as selection based on a test, or assignment to different training or working

conditions), most validity studies linking a psychological trait to an occupational outcome are correlational. Causality is commonly assumed from temporal order, but strong evidence for causation is rare.

Correlational data are subject to uncontrolled variance. Statistical techniques may be used to correct for obvious sources, but not all sources are obvious. These conditions present challenges for modeling, not the least of which is that the presence of noise on at least the order of the effect size can obscure the effect in any visual evaluation.

As in many fields of psychology, historically, small sample sizes have been more typical than large ones. Data gathering was effortful and costly. It required the cooperation of employees and managers in assigning time to processes which were not relevant to operating a business, such as filling out surveys. Data sets in the hundreds or thousands of cases were available, but even those might not have the power to detect small effect sizes. In recent decades, meta-analyses were able to extract results from these smaller studies.

Compounding the problem of uncontrolled variance, the available criterion measures were, and indeed still are, often poor representations of the variable of interest, such as secondhand or retrospective reports (Steinberg et al, 2000).

Recently, large-scale warehousing of business data has become feasible. This has led to "data-mining" operations in numerous fields of study, in which data collected for the purpose of business are sifted through for theoretically interesting relationships. Marketing research, for example, may compare purchasing profiles of different

demographic groups, or link the frequency of one type of purchase to the frequency of another. Datasets of this type may have cases in the millions, if one case is a person.

Data mining is typically exploratory, and has sometimes been written off as a form of dustbowl empiricism. The practical utility of a relationship may, for example, lead to the acceptance of an ad hoc theory. On the other hand, by the nature of exploratory analysis, relationships may be discovered which were not expected, or which were too subtle to detect in smaller traditional studies. Confirmatory studies, such as determining the predictive validity of an assessment, also benefit from the larger sample sizes.

Despite the availability of large samples, there remain intrinsic problems with the type of data available. The criterion is no less vulnerable than the predictor. In fact, it may be more so. The variable of interest may be as broadly defined as "job performance" or "incidence of counterproductive behavior", or may be much more specific. However, these quantities are often not measured at all, and if they are, are subject to flaws that lower their reliability (Steinberg et al, 2000; Scarborough, 2005).

Managers' evaluations of employees are subject to the influences of irrelevant factors (e.g. personality factors on an ability judgment), halo effects, leniency, severity, and central tendency. There may be implied incentives in place for good reports. On the other hand, the average incumbent employee is probably better than the average candidate, and so their scores may be lowered by comparison with available examples. Empirical performance records such as cash register speed or sales volume may be compromised by low compliance, as well as effects of time of day, season, and co-worker

performance. Even hire and termination records may be incomplete or inaccurate due to manager noncompliance (with corporate rules, in this case) or administrative delays.

Restriction of range is a further problem which is not corrected by sheer sample size. If a valid test is used for selection, its apparent correlation with criteria measured only on the selected population will drop. There are statistical corrections for this effect (Lord & Novick, 1968), but they are dependent on several assumptions which are often violated in practice, and others which are difficult to check. When possible, it is best to "try out" a test on an applicant population and validate it before it is used to select anyone; on the other hand, even this procedure is compromised if any selection process is in use which correlates with the outcome of the test. A different test may be such a process, but so may the informal judgment made by a hiring manager (Autor & Scarborough, 2004). Because the uncorrected validity coefficients are conservative, they may be considered a minimum for realized validity.

It may be considered a benefit of large-scale automated standardized assessment that it is easy to detect subtle effects of applicant characteristics. For example, thousands of cases give plenty of power to test for discrimination against protected groups, or even differential item or test functioning. Regional differences are apparent; even site-to-site differences within a city are relevant. However, the proliferation of such findings is also an indication of overall data quality. Unless given meaning in terms of psychological constructs, these incidental findings obscure the relationship between assessment score and outcome.

No efforts to reduce extraneous, measurement-induced variation in the predictor or criterion data will make the model fit well if the test is based on the wrong psychological model. Researchers always run the risk of this, but have compounded the problem by putting all the eggs in one basket. Overwhelmingly, researchers relating personality to occupational performance have tested linear models. The reasons for selecting a linear model include simplicity, comprehensibility, ease of computation and relatively low sample size requirements. A linear model can be easily translated into a test scoring algorithm, possibly involving weighted sections. Some psychological theories specify a linear or proportional relationship for stronger reasons, but others do not. In order to account for more of the variation among employees, it may be necessary to adopt nonlinear statistical models and more complex modes of scoring tests.

1.2. Neural Networks

One type of mathematical model worth considering is the artificial neural network. Inspired by the behavior of nerve cells, neural networks perform distributed computations across numerous "nodes". Neural networks are not just a model of human cognition; they can be used as a general statistical model to predict an outcome or set of outcomes from a set of inputs.

Artificial neural networks are computationally intensive, but well within the capacity of cheap modern computers. They are also adaptable to a wider range of actual functional relationships between independent and dependent variables than classical

statistical techniques in the industrial psychologist's toolkit, such as linear multiple regression. They are able to systematically "learn" directly from data in the absence of extensive human interpretation. They do not require, for example, that the salient interaction effects be pointed out to them beforehand.

We shall not concern ourselves greatly with the similarities between artificial and biological neural networks. One reason for this is that there are many differences which place limits on the obvious analogies. Artificial neural networks have developed utility without being a high fidelity representation of a biological neural network such as the human brain. They are typically much less complex, having on the order of 10^3 units rather than 10^11 (Scarborough, 2005). Computation of each unit occurs on a different time scale, by a factor of as much as 10^6. The function of each artificial node is restricted by comparison to the possible processes carried out by the corresponding biological neuron. Nodes record inputs and transmit outputs in a synchronized manner, whereas neurons fire asynchronously (Hertz, Krogh & Palmer, 1991). Finally, the brain learns by chemical changes that occur when sequentially linked neurons fire together, a process known as long-term potentiation. While there are neural network learning methods, such as the Hebb rule, which approximate this, other methods such as backpropagation do not (Crick, 1989).

A second reason to dismiss consideration of biological neural networks is that the human brain is not the ideal toward which we would have our computer programs aspire (cf., Hunt, 2002). Artificial neural networks here are used in the context of employment outcome prediction. The perfect employment test is probably not a simulation of a human

being briefly reviewing a job application. Human beings make errors, fail to perceive large patterns, fail to perceive salient details, and exhibit systematic biases. The purpose of employment testing is to alleviate these problems, not to replicate them.

When used in their capacity to model statistical patterns, rather than emulate human thinking about patterns, artificial neural networks (henceforth "neural networks") can be of use to industrial psychology. Thus far, use has been limited, but reasonably successful. Garson (1998) notes, "Neural networks may outperform traditional statistical procedures where problems are unstructured, involve incomplete information, are ambiguous, and involve large sets of competing inputs and constraints, provided the researcher can accept approximate solutions." This appears to be a fair description of the social sciences. Further, two primary requirements of neural networks, namely large data sets and fast computers, are no longer difficult to find.

Neural networks in industrial and organizational psychology usually operate in two modes: classification and prediction. Elsewhere in science and engineering they are also used for pattern completion, control, and constraint satisfaction (Garson, 1998), but these uses have not appeared in the domain of organizational psychology.

Classification is of use for some organizational applications. For example, Somers (2000) used a self-organizing map to categorize employees in a hospital setting into four groups based on measures of organizational commitment. Follow-ups showed different patterns of behavior between these groups, but the modeling took place prior to measurement of the outcome variables and was descriptive in nature. Such exploratory

contexts are ideal for clustering and classification techniques. This paper, however, focuses on prediction.

A neural network operating in this mode may predict either continuous or discrete variables. The latter form may also be called classification, in the sense that the neural net is learning an existing categorization, but this is not to be confused with the classification methods described above. Unlike those methods, the neural network does not invent a classification according to the structure of the inputs, but rather attempts to describe the structure of the outputs in terms of the inputs.

In this context, traditional alternatives to neural networks include discriminant analysis and linear regression (Garson, 1998, pp. 81-82). Both of these techniques can be defined as neural nets on which restrictions have been imposed, special cases, but they have advantages related to their simplicity. They have been extensively studied and are well known. Their parameters are computed explicitly in a single step using linear algebra. Both the models and the resulting parameters are easily explained.

On the other hand, unrestricted neural networks better describe nonlinear relationships and interactions and may thus explain more criterion variance. This has been demonstrated repeatedly, including in organizational research. For example, biodata or personality variables appear to predict turnover better when the method used is a neural network than when multiple linear or logistic regression are used (Dempsey et al, 1995; Somers, 1999). Further, neural networks are more robust than linear discriminant analysis where data may be missing, a common condition in industrial psychology (Collins & Clark, 1993).

Neural networks address a need for arbitrary nonlinear multivariate modeling in organizational contexts, as well as in other areas of psychology. The reason this need exists can be explained with two propositions. One proposition is that not all relationships between meaningful psychological measurements are linear in nature. The second proposition is that because linear methods have been readily available, those relationships which can be described well by a line or plane are likely to have already been studied and described, compared to those which cannot. The set of linear true relationships has been tapped into by investigation, and the set of nonlinear true relationships has barely been touched.

When should a researcher consider linear modeling to have failed? When low effect sizes and lack of significance occur, the usual suspects are various forms of measurement error, including poor reliability of measures, and the moderating effects of additional variables. However, a weight of accumulating evidence, such as repeated fruitless efforts to improve measurement, may indicate a misspecified model. When the components of the model make both theoretical and "common" sense, the next suspect is the mathematical form of the model (Scarborough, 2005). Further evidence may come from residual plots and other visual diagnostics, but the relationship may not be easily perceived because of its still-small effect size, or it may require multiple predictor dimensions.

As an example in organizational psychology, consider job satisfaction and job performance. It is intuitively obvious that the two should be related, and yet many studies have failed to find a clear relationship. One recent study found a nonlinear relationship

between those two variables and either role conflict or job involvement. In the space defined by role conflict and job satisfaction, or job involvement and job satisfaction, there were regions in which the effect of job satisfaction on job performance was strong -- very nearly a step function. In other areas, however, there was little effect of small changes in either predictor variable on job performance. In this case, measuring a variable such as job satisfaction across a wide range, or over the wrong narrow range, would lead to a lowered slope in a linear fit (Somers, 2001). Under the assumptions of the linear model, it is irrelevant whether the experimenter measures the right range of a given variable, so a solution leading to more consistent and theoretically sensible effect sizes was not apparent.

Scarborough (2005) recommends that the assumption of linearity, inherent to most psychological studies, be subject to empirical test. Such a test would evaluate the fit of the linear model by comparing it to an arbitrary nonlinear model such as the neural net, rather than being an error-prone visual assessment conducted by the experimenter.

For the problem at hand, it is convenient that a neural network will model either a linear relationship or a nonlinear relationship equally well. The form of the model is not as important as the quality of the resulting predictions. It is possible that in predicting a given employment outcome, even a neural network will discover only linear relationships, and a linear regression model would predict the outcome just as well. Experience suggests it is likely, however, that at least one of the variables has a region of particular sensitivity, an optimal point, or a non-additive interaction with another. Therefore, the more flexible model, the neural network, will be used.

1.2.1. Network architectures

There are several architectures under which neural networks may be constructed. Not all of them are discussed here. Specifically, the architectures can be divided into two broad classes based on the type of problem which they are designed to solve, and the type of training they undergo.

The first type includes networks that produce feature maps, clusters, and other descriptions of the data without reference to a criterion. They are trained by unsupervised learning, that is, also without reference to a criterion. While useful for some purposes, such as the organizational commitment study mentioned above, these networks do not meet our needs when, prior to hiring an employee, we wish to predict outcomes on the job.

The second type are trained to predict a criterion, using examples where the criterion as well as the predictors have been measured. This process is known as supervised learning, because it requires a "supervisor" to check the network's prediction for each case at each step of training and send back a description of errors made. The parameters of the network are then adjusted to reduce the error. In this way, the network's predictions are tuned to the data.

Supervised learning may be considered a one-step form of pattern recognition, as opposed to the classical two-step form in which feature extraction precedes prediction according to features (Haykin, 1999). Other than behaviorists who treat the brain as a

"black box," psychologists typically use the second form; we first define constructs, and second develop a theory of how those constructs lead to observed behavior. Neural networks do not require the specification of meaningful constructs. Multilayer networks do perform an additional step of feature extraction beyond that involved in measuring the inputs, but the only labelling of the features is the equation relating them to the criterion.

Not all architectures within this category are useful for our purpose, but many are. One limitation on the architectures is that they must be feed forward networks. That is, information flows in only one direction (excluding error data during training), from the inputs toward the outputs. The alternative is a recurrent architecture, which has one or more loops internally, such that internal components of the network may contribute to their own states. A recurrent network thus has a "memory" for one or more previous rounds of calculation. In addition to being mathematically intractable for certain purposes, such a memory is not a desirable property in the context of employee selection. It is not fair to a job applicant to base the hiring decision partly on one or more previous applicants, without the possibility of reciprocation. The fact that exactly this kind of comparison may occur in interview situations is not relevant; it should not be emulated.

There are several types of feed-forward architecture. We will consider only one example, the multilayer perceptron, but the results generalize.

The perceptron is a classic form of neural network, and the multilayer perceptron is a homogenous evolution. It is relatively transparent mathematically.

The multilayer perceptron is composed, as its name implies, of layers of nodes. Each node is an identical functional unit, described below, which accepts inputs and

produces an output. The outputs from the nodes on one layer are the inputs to the nodes on the next layer.

There are at least three layers of nodes in the multilayer perceptron; the classic perceptron had only two, input and output. Input nodes are those that represent quantities extrinsic to the network; output nodes are those that produce the neural network's responses. The multilayer perceptron has additional layers between the inputs and outputs, and no direct connections from input to output. These in-between layers are called hidden layers. Their states are not typically meaningful in a concrete sense, and they are generally not reported, but they greatly increase the modeling power and therefore usefulness of the network.

Minsky and Papert (1969) noted that the classic perceptrons, lacking hidden layers, could only distinguish linearly separable sets. This is a severe limitation in terms of real-world modeling. Not only must the right information be chosen, it must be presented in the right form, be that a ratio, a power of an observed quantity, or some other transformation. Consider, for example, the set of points within a radius r of some center and those which are outside r, with each point given as a coordinate pair to two inputs. Although the condition is simple, a perceptron could not approximate it to any great precision. However, in cases such as this where the sets are nonlinearly separable, the presence of a hidden layer can allow for an arbitrarily adjusted nonlinear transformation into an alternate space where the sets are linearly separable -- for our example, some arbitrarily good approximation of radius-angle space.

Theoretically, only one hidden layer is required for even the most complex relationships. Additional layers sometimes provide a more parsimonious or understandable explanation, however. This is most justifiable when the researcher knows a priori that there are higher-order relationships present in the data. It is rare to see more than three hidden layers in use. (Garson, 1998)

The default configuration of a multilayer network is to have each node in a given layer receive for its inputs the states of all the nodes in the previous layer. This is known as being "fully connected". However, if the researcher knows something about an overarching structure connecting the inputs, some connections may be "pruned". This means that the receiving node only accounts for information from some of the nodes in the previous layer. If it is possible to prune a network from a priori knowledge, it is advisable to do so, as it cuts down on noise. (Haykin, 1999)

The method described in this paper imposes what structure is known prior to transmission of any data to the neural network. Therefore, pruning is not practical in this case.

The structure of each node is identical, and can be described by the equation:

$$\text{output} = f(\ \mathbf{weights} \bullet \mathbf{inputs}\ ) \qquad (1)$$

where weights and inputs are vectors of equal length, and output is a scalar quantity.

The node is usually represented diagrammatically with two parts, as in Figure 1. The first part is a summation. Specifically, it is a weighted sum of the inputs to the node,

represented by the dot product of vectors in the equation above. There is exactly one input which does not come from a previous layer; it is always set to unity, and the weight by which it is multiplied is known as the bias.

The second part is the transfer function, $f()$, which scales and transforms the weighted sum into an output. In the simplest case, the transfer function is linear: $f(x)=ax+b$. In this case, the computation of the multilayer perceptron can be reduced to matrix algebra and cannot model nonlinear relations between variables (Bishop, 1995).

A common transfer function is the step function, set equal to 1 above a threshold value and 0 (or -1) below it. This is the classic transfer function, and may be implied by the use of the term "perceptron"; some recent authors use the term more liberally. Several variations on the binary step function exist, including trinary step functions which report

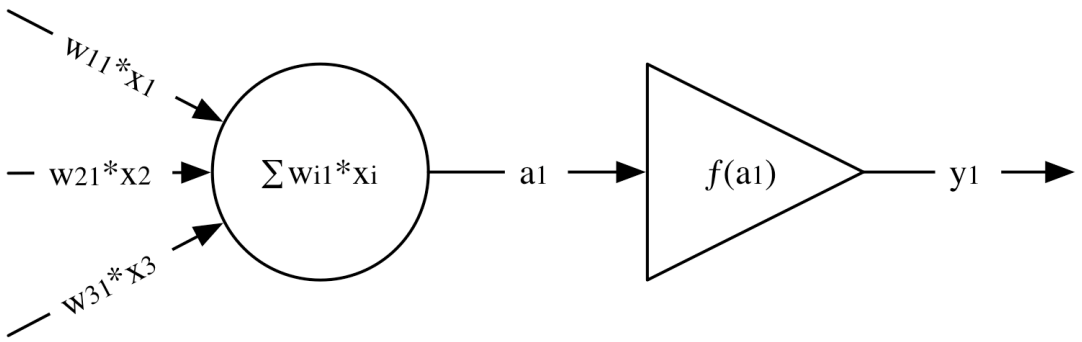

Figure 1. One node of a perceptron.

0 at the threshold, 1 above and -1 below. Clipped linear functions restrict output values to a specific range while maintaining linearity.

The transfer function need not be monotonic. In some cases, Gaussian distributions are used. These are localizing functions, which essentially report whether the sum of inputs falls within a particular range.

For our purposes, the most interesting are a set of functions that are smooth, differentiable, and monotonic. This class of functions, the sigmoids, is commonly used. It includes the normal ogive, otherwise known as the cumulative normal distribution, although that is uncommon for reasons of mathematical tractability. The logistic function, when compressed horizontally by a factor of 1.7, falls within 0.01 of the normal ogive at all points (Birnbaum, 1968) and is for practical purposes equivalent. A third function, the hyperbolic tangent function, is a further rescaling and vertical shifting of the logistic, in order that it should range from -1 to 1 instead of 0 to 1 and be antisymmetric around 0 (Haykin, 1999). This improves the speed and probability of success of the training process (Bishop, 1995).

The multilayer perceptron is one example of a continuous function estimator. Provided that it has at least one hidden layer with a nonlinear transfer function, and provided sufficient nodes and training cases, a multilayer perceptron can approximate any continuous function arbitrarily precisely. This can be shown by the universal approximation theorem (Haykin, 1999). In practice, one is typically more concerned with overfitting the training data set, including modeling error, than with having too few

parameters to fit the real variation. Overfit leads to poor generalization to future data points which have errors independent of any of the training cases.

In light of their ability to model arbitrary continuous function surfaces, three-layer perceptrons are excellent for predicting near-continuous data such as revenue per hour, as well as job tenure, dollar amount of theft, and other business metrics.

To predict qualitative or otherwise non-continuous data, one may divide the cases at a threshold output level. This results in a classification. (Haykin, 1999, p. 185) If there are more than two categories, the network can be trained to produce a separate output for the probability of membership in each possible category. This can be used, for example, in the prediction of separation reason. However, there are more efficient ways to go about it, which may result in better predictions. A multilayer perceptron may have more than one output, giving a probability of membership in each category. Similarly, several networks may be trained, one for each category; this, however, allows the possibility of two categories being predicted. Finally, other network architectures may be better suited to categorical prediction.

1.2.2. Useful Properties

There are several properties of neural networks which will be of use in adaptive input selection. These properties are not specific to the multilayer perceptron or to the radial basis function, but apply at least across the entire class of feed forward networks which are trained by supervised learning.

In devising an algorithm to feed information adaptively to a neural network, we will be concerned with error of prediction. Specifically, we will be concerned with changes in the amount of error. Fortunately, the problem of describing the errors the network commits arises naturally in the context of training the neural network. Optimizing predictive accuracy requires a means of describing the errors the network commits in predicting the training cases. Typically, a scalar error function is minimized by a variety of methods. These methods refer to a "performance surface", where the error quantity is treated as a function of the adjustable parameters of the network. In the case of the multilayer perceptron, the parameters are the weights, including the biases, entering each node. In the case of the radial basis function, the parameters also include radii and centers of the hidden nodes.

The error function is usually the sum of squared differences between the actual levels of the outcome variable and the corresponding predicted levels in all the training cases. Variations include the mean squared difference. The choice of this function was based on the assumption that errors will be distributed normally, but the use of the least squares method does not require that assumption. According to the Gauss-Markov Theorem, the only requirements are that the errors be independent and identically distributed with finite mean and variance (Neter et al, 1996). Several alternative performance measures have been suggested, including entropy (see, e.g. Bishop, 1995).

Neural networks have the property of graceful degradation in the presence of erroneous data. In the general case, this only means that the functions they fit are continuous and thus that small perturbations of inputs result in small perturbations of

outputs. However, if a bounded transfer function is used between layers, the neural network will still give a similar output even if one or more inputs is replaced with an extreme or nonsensical value. This is particularly valuable in mechanical applications (Haykin, 1999), but is also useful for our purposes.

In all cases, it is assumed that there is a value for each input. That may mean that a default value is substituted for missing data, or that a random or erroneous value is expected. What is important here is that regardless of the value of any given input, the other inputs still meaningfully restrict the possible range of the output. The uncertainty of the output value decreases monotonically with each input which is known to be valid. It also decreases monotonically with the uncertainty of each input, so that if one input is restricted to a subset of all possible values, the output is restricted as well.

In typical applications of neural networks, missing data is not intentional on the part of the developer, and values which are not missing (or which are substituted for missing data) are considered exact. The missing data may be accomodated either as unsystematic, through the network's general robustness, or as a systematic indicator of a failure condition. In the latter case, the missing data code is a relevant value in itself, if it is available. Unsystematic substitutions for missing data may not result in a distinct code, but a random value. This happens, for example, in mechanical systems where input-generating components may be susceptible to analog "noise", or in electronic network communications where single-bit errors may be introduced. This type of substitution is less diagnostic; the network only knows there is an error if the value violates the expected

relationship between inputs. Even then, it may only be possible to tell that an error is present, not identify which input gave the bad value.

Uncertainty about measured values due to measurement error is typically either not accomodated, or implicitly accomodated by the training set. In mechanical applications, the error of a particular instrument is likely to be constant over time. It simply increases the unaccounted-for variation after the relationship between input and desired output is measured.

In this application, inputs will always be missing by design, although the training set has no missing data. Further, some measurements which are entered as inputs have error quantities which change over time and which are large enough to change the output. A numerical method for estimating the effect of incremental uncertainty in the inputs on uncertainty in the output will be discussed.

Another quantity that will be useful is the sensitivity of an output to an input. This is the amount of variation in the prediction that results from small perturbations in a given input. If a nonlinear transfer function is used, this sensitivity will vary across the values of each input, including but not limited to the input for which it is being calculated. For that reason, it is calculated as a partial derivative of the output with respect to the input, with all other variables left in the equation (Bishop, 1995; Montaño & Palmer, 2003).

1.3. Computerized Adaptive Testing

A recent trend in testing has been the transition from paper-and-pencil to computerized adaptive testing.

By definition, a computerized adaptive test (CAT) is any test which meets two criteria. The test must be administered by a computer, making it computerized. Further, over the course of the test, the examinee's performance influences the items presented. In practice, the term computerized adaptive testing has become associated with a much more specific concept, a form of computerized adaptive test which estimates a unidimensional latent trait according to the principles of item response theory. Over the last three decades, CATs of this more specific form have been found to work well and have been implemented in large-scale applications. This paper describes a CAT which does not adhere to this form, but which meets the definition.

Computerized tests need not be adaptive, and adaptive tests need not be computerized. Individually administered tests may have rules regarding progression to more difficult items or discontinuation of a subtest (e.g. classical digit span tests which stop when two consecutive errors are made), or they may be informally adaptive, as in the case of a teacher's diagnostic questioning of a student. No computer need be involved if a human being is there to select the next item.

Mass-administered paper-and-pencil tests are typically not adaptive. A paper-and-pencil test, the self-scoring flexilevel test (Lord, 1971), has been proposed which gives a harder item following a right answer and an easier item following a wrong answer,

provided the examinee correctly follows the choose-your-own-adventure-like cues for which item to answer next. There are technical complexities involved in printing such a test, however, and examinees can easily render their tests unscoreable by errors in next-item selection. These issues are alleviated by computer administration.

Large-scale application of computerized adaptive testing was made possible by the mass production of the microcomputer. Segall and Moreno (1999) report that early attempts to administer CAT using mainframes were largely unsuccessful; the test taker could not count on a prompt response when other users were on the system. Some diagnostic and formative classroom applications achieved success (e.g. Braunfeld, 1964); the difference may be smaller numbers of test takers, or that Segall and Moreno refer to the narrower definition of CAT.

The United States military initially expected the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) to require custom-built computers, in order to satisfy the requirements of portability, processor speed, and high (for the time) graphics capability. But by the time the project was underway, IBM-compatible personal computers and early versions of Windows were satisfactory (Segall & Moreno, 1999). The technology available has fast exceeded the demands of testing.

Adaptive testing was an early use of computers in testing, and many authors (see, e.g., Wainer, 2000) treat the costs and benefits of adaptive testing as being inextricable from those of computer presentation of a test. One book even goes so far as to refer to non-adaptive, non-multimedia computer presentations as "page-turners" for tests equivalent to their paper-and-pencil forms (Drasgow & Olson-Buchanan, 1999). But with

the more recent rise in networking, testing programs have begun to computerize for reasons of cheap, instantaneous transmission and score reporting, rapid updating and recalibration, and centralized storage of data. Altering the nature of the test to an adaptive form is today a separate consideration from simply putting the test online.

Computerized adaptive testing has been used in numerous educational applications, including assessments of both academic aptitude and achievement. It has been applied to assessments as varied as a test of tonal memory, which is a component of musical aptitude (Vispoel, 1999), a form of the NEO-PI-R (ref Reise & Henson, 2000), and an evaluation of headache impact (Ware et al, 2003). Because the development of a CAT is involved and costly, most of the non-research applications have been large-scale projects, however, measuring examinees numbering at least in the thousands.

Employment testing applications of CAT have included the CAT-ASVAB as well as several certification examinations (see, e.g., Bergstrom & Lunz, 1999). The online testing company Brainbench (http://www.brainbench.com/) exclusively uses adaptive forms of its numerous skills and certification tests. Zickar et al (1999) developed a computerized adaptive math test for applicants for a State Farm programmer-analyst position, joining it with a straightforward collection of biodata; State Farm's main concern was test security, justifying the expense of CAT for a small population.

Adaptive testing has several advantages over conventional testing, particularly when computers ease the computational burden. These advantages are above and beyond those conferred by computer administration.

First, CAT allows more even measurement across the entire range of a trait. A conventional ability or skill test, for example, typically contains items that are easy, moderate and difficult. Almost all the items provide information about an examinee of moderate ability. However, an examinee of high ability who demonstrates proficiency on the moderate items can be expected to answer the easy items right; they provide no additional information because they have zero variance. Similarly, an examinee of low ability can do nothing more than guess wildly at difficult items, adding noise to any estimate of their ability. The result is that the standard error of measurement is not constant across the range of ability, as classical test theory would suggest. Error is inflated and reliability is decreased for high or low ability examinees.

CAT uses early items to target the difficulty of later items. An examinee who shows proficiency early on will receive more difficult items than one who answers the first few items incorrectly. This means that examinees at either end of the ability range answer few non-informative items, and more informative items. These "extra" hard or easy items reduce the standard error of measurement in the high and low ability ranges. The CAT is still not likely to produce exactly the same standard error of measurement in the same number of items for every examinee, but it will be closer to that ideal than the conventional test. (Thissen, 2000)

These effects are not limited to ability; an analogy can be made to any unidimensional construct. Ability is convenient in that the terminology is familiar.

By the same mechanism, adaptive testing is faster than fixed-sequence testing for the same precision of measurement. Even the paper flexilevel test halved the length of

administration for ostensibly the same depth of measurement. Computerized tests, given a variety of items, may achieve excellent performance after asking a small number of questions. (Mislevy & Bock, 1989)

These advantages are in areas affected little, or not at all, by neural network prediction. Neural network prediction does reduce the effect of irrelevant items in scoring, but it does not entirely eliminate them from administration. It is reasonable to say that it would be a disadvantage of neural network prediction if it could not be used with adaptive testing, and vice versa.

In order to consider the technical issues involved in using CAT in conjunction with neural network scoring, it is necessary to understand the mechanics of a traditional CAT. Components may then be systematically replaced, without changing the broad principles of operation. There are two components that will be of particular interest. One is the item selection algorithm, according to which the next item is chosen. The other is the scoring rule, a mathematical procedure according to which the examinee's item responses are converted to a score. If the scoring rule is a neural net, how does the item selection algorithm need to change?

The traditional form of a CAT is an assessment devised to measure a unidimensional construct such as (but not limited to) ability. The principles of item response theory may be applied to both item selection and examinee scoring.

The test measures a single latent trait, on which the examinee's true score is θ. An approximation of θ, $\hat{\theta}$, is available at any given time; $\hat{\theta}$ is used to select the next item according to its difficulty (and possibly other parameters). A convenient feature of item

response theory is that the item and the examinee may be placed on the same scale. An informative item is therefore one whose information function is high in the neighborhood of $\hat{\theta}$. The information function is defined as the derivative of the probability of a keyed response with respect to θ, and therefore it can also be said that an informative item is one for which a small difference in the latent trait makes a large difference in observed response. In the simple case of items which conform to a one-parameter logistic model, the most informative item is the one whose "difficulty" most closely matches $\hat{\theta}$.

The preceding description has been vague in order to accomodate several similar scoring rules, each of which correspond to a slightly different item selection algorithm. One school of thought is descended from the practice of maximum likelihood estimation; another traces its roots to Bayesian inference. The primary difference, not affected by technological capabilities, is whether $\hat{\theta}$ should be calculated conservatively according to assumed population parameters, or purely according to the examinee's responses. Recently, the estimator most often used for θ is the expectation a posteriori (EAP) value, which unlike the maximum likelihood value is robust to bimodality and other distributional anomalies that may arise. (Bock & Mislevy, 1982; Thissen & Mislevy, 2000)

In any case, once the item is selected and responded to, the distribution from which the examinee is assumed to come is updated according to the scoring rule. At first, the examinee is assumed to come from the distribution of all examinees, which may be constant (as in the case of maximum likelihood estimation), normal with zero mean and unit standard deviation, or an arbitrary distribution corresponding to a known population

subset. After one item, the examinee is assumed to come from the distribution of all examinees who made one particular response to that item. After the second item, the distribution is restricted by two responses, and so on. The process of updating from one distribution to the next amounts to a convolution of the existing distribution with the characteristic curve for the given response, where the characteristic curve is the function relating θ to the probability of giving that response. (Wainer & Mislevy, 2000) $\hat{\theta}$ is recalculated from the new (posterior) distribution; in the case of the EAP, it is the mean.

In all variations of the traditional form of CAT, the scoring rule and the item selection algorithm are intertwined with and optimized to each other. In order to use a scoring rule which is not based on item response theory, an item selection algorithm must be devised to match it. Not all scoring rules have the mathematical conveniences of item response theory, such as the examinee and the item being on the same scale. However, functional equivalence is possible.

There is one additional variation on the scoring rule that has been extensively studied, including in employment contexts, and which is worth some attention. Computerized adaptive testing is occasionally applied to situations in which only a pass-fail judgment is required, not a relative score which may be compared to other examinees. This may well be the case in an employment setting, where the test may be used as an early screening, followed by more intensive evaluation. However, if the cutoff score is known in advance, it is more efficient to target the items to maximally discriminate at the cutoff level, not at the examinee's probable ability level. The cutoff never changes, so there is no reason to make the test adaptive at all. (Wiberg, 2003;

Zickar et al, 1999) If additional information may be useful, but there is a threshold value which is important, Bergstrom & Lunz (1999) call for a CAT with an item pool distributed such that most of the items measure near the threshold. That way, it is still possible to identify an outstanding candidate, but ones who are near the threshold are measured with a high degree of precision. It is not necessary to know how far below the threshold a candidate falls, merely to be certain that the candidate did fall below threshold.

Wiberg (2003) and Bergstrom & Lunz (1999) are concerned with mastery testing, where the cutoff score is relatively permanent, and thus do not discuss what to do when the threshold is subject to revision after the item pool is fixed, a situation that may come up in employment contexts. If an employer may lower or raise the threshold depending on the availability of job applicants during a given time period, then targeting the entire item pool to the cutoff score is shortsighted. Targeting a given test, however, may be a viable option.

The cutoff argument, while presented as unidimensional in the context of mastery testing, may be generalized to the prediction of category membership in multiple dimensions. In general, it is advisable to consider whether there are regions of latent trait space where information is more valuable; otherwise, one implicitly assumes equal value throughout that space.

1.4. Subtests and Scoring Considerations

A major difference between the tests typically converted to computerized adaptive form and assessments of personality in the prediction of employment outcomes is that the latter are typically not unidimensional. Recall that job performance and job tenure are composite criteria, influenced by several variables. An assessment may involve several corresponding variables, particularly if biodata are used.

In scoring such a multidimensional test, it is useful to know what dimensions are being measured. This is not only for the purpose of interpretation; it anticipates the need for diagnosis when, for example, a social change leads to the erosion of validity. If interpretation is to be done, the theoretical expectation that certain items will measure certain constructs must be verified empirically. When the dimensional structure of the assessment is understood, unidimensional subscales may be constructed such that they exhibit internal consistency.

The use of subscales both complicates and simplifies our later task. From the perspective of a neural net, a well-constructed scale reduces largely redundant information to a single estimate with less noise. This reduces the number of training cases needed and may improve performance, because the data points are located in a lower-dimensional space (Bishop, 1995). However, the trait estimate produced by a subscale is qualitatively different from a direct representation of an item; it is continuous and comes with an uncertainty, whereas an item response is categorical and concrete. Either the applicant chose "1" or he did not. For this reason and the length of application, a subscale

requires differential treatment by the selection algorithm to be developed. Nevertheless, efficiency of training outweighs elegance of the selection algorithm. We will use subscales where it is possible and reasonable.

It may seem evident that the most straightforward way to determine the dimensionality of a set of items is through factor analysis. Factor analysis is, however, only one of several methods. It may not be the most appropriate method for item-level personality data. Factor analysis assumes the items are continuous, and many of its significance tests further assume the responses are normally distributed, but a more likely case is that each item has only a few discrete possible responses. This case leads to underestimated loadings and overestimates of the number of factors present (Embretson & Reise, 2000, p. 308). It is also subject to a form of indeterminacy which is likely in this type of application. Doublet factors, or constructs which are represented only by two items and which are not correlated with other factors, can result in improper solutions (negative variances) or solutions which do not accurately reproduce the underlying structure (McDonald, 1999, p. 180-181), and thus cannot be expected to replicate in independent datasets.

Wanek, Sackett and Ones (1998) used an alternative method to explore the dimensionality of seven integrity tests. First, two industrial psychologists independently sorted the test questions into groups by content and attempted to name each group. They compared the group names resulting and settled on nomenclature, then came to a consensus about item placement, entirely without reference to examinee data. Finally,

reliability was calculated for each resulting subscale and items with intraitem correlations consistently below 0.1 were dropped.

Variations on the exact method can be easily imagined. The significance, however, is that empirical exploratory methods may be entirely bypassed when the theory linking item content is strong. It is also worth noting that Wanek et al did not bypass the confirmatory evaluation of internal consistency, nor further assessments of convergent validity. Those confirmatory evaluations were considered valuable, even when the exploratory analyses were not.

When criterion data is available, a third method may be used, as demonstrated in the development of the California Personality Inventory. "...[D]evelopment of the CPI made no reference to factor analysis. Instead, the method of criterion-keying was used: items were chosen on the basis of their ability to discriminate criterion groups." (Matthews & Deary, 1998; p. 23)

This method is unconventional in psychology, where construct validity may be favored over criterion validity. Criterion-keyed traits may disagree with those which are gleaned from factor analysis, and may or may not achieve high reliability. The Occupational Personality Questionnaire did, but some tests which predict occupational outcomes may do so by predicting several intermediate behaviors which all contribute to that outcome. (Matthews & Deary, 1998; Hunt & Paajanen, 2003)

Cluster analysis is another set of methods related to factor analysis. Items can be clustered according to correspondence across individuals. Methods such as agglomerative

nesting may produce a useful atheoretical guide toward linking items. As with criterion-keying and content-based sorting, empirical validation is still called for.

Any of the four methods described above can be used in conjunction with each other to provide converging evidence for the dimensional structure of a test. All but factor analysis were used in the development of the subscale structure covering the majority of the assessment used in this investigation. Final decisions about inclusion and exclusion of items were made on the basis of incremental reliability and expert judgment regarding content. An example where expert judgment overrode reliability involved the high correlation of a risk-taking item with several sociability items in a population of athletes. The correlation was not expected to generalize.

Provided that each scale is defined without distinguishable subsets of items which are more intercorrelated, constituting a local independence violation, the subscales can be assumed to correspond in a one-to-one fashion with latent traits of the examinee. (Lord & Novick, 1968; cf., McDonald, 1999, p. 257) This is in contrast with the entirety of the assessment, which predicts a single employment outcome but contains more tightly-coupled scales within itself. Thus, for each subscale, a latent trait (or item response theory) model may be applied to its items.

A number of researchers have suggested extensions of IRT models to multidimensional tests. These methods allow each item to provide what information it has available to the estimate of the examinee's placement on each dimension, in contrast to having several independent measures of the different dimensions. Muraki and Carlson (1995) developed a form of factor analysis which assumes that polytomous items call for

a linear combination of several latent traits. That is, each item has a "direction of measurement" vector in a space defined by several traits, and can be described by a one-dimensional curve along that vector. Embretson and Reise (2000) discuss a "non-compensatory" model in which several abilities are required to solve a problem. In contrast to Muraki and Carlson's model, the non-compensatory model does not predict that an examinee high on one quality can make up for a low score on another. This model cannot be described by a one-dimensional curve along a "direction of measurement" regardless of perpendicular position. We will not use either of these methods, however, because the neural net is capable of the representing such a model, while less intrinsically constrained.

The latent trait model focuses on shared variance among a set of items. That shared variance is considered to be the best measure of the underlying trait. Sum scores and more complex trait estimates discard unique variance which is not common to the set of items as a whole. This has two consequences.

First, the reduction of a set of items to a superior measure of their shared variance is the reason that a trait estimate can be used as a form of compression of the item responses. If the latent trait is what predicts the outcome, then unique variance of each item is just noise. The principle of local independence implies that the noise is random and will, on average, cancel out.

Second, the removal of unique variance may remove useful variance. Citing the multidimensional nature of job performance, Hunt and Paajanen (2003) advocated heterogeny in the test as a whole, including shorter and less internally consistent scales,

in order to better sample the range of personality traits affecting a performance measure. Further, it is possible that an item response may be driven by both a trait which other items also measure, and a second trait which is linked to the criterion but not measured by other items.

In order to preserve useful unique variance, as well as justify the assumption of local independence, items which appeared to be internally complex or which did not link strongly to scales were scored individually, not entered into scales.

2. A hybrid selection algorithm

Up to this point, this paper has described prior research and background information. From this point forward, I will present my dissertation research in which I have developed a combination of neural network modeling and adaptive testing. Although each of these approaches has been researched separately, the combination of the two is to the best of my knowledge completely new. This chapter describes in mathematical terms how to combine neural network modeling and adaptive testing.

Previous computerized adaptive tests have depended on item response theory for parameter selection and to guide item selection. Previous neural networks used in employee selection have assumed that all input data is present, or is missing completely at random. In order to reap the advantages of both adaptive testing and neural network scoring, a new set of rules are needed to govern which items are presented and omitted, and to interpret the output of a neural network whose input data is missing in ways constrained by present data. One major purpose of this paper is to propose such rules; the other is to show that they work as designed.

The following section develops an adaptive selection algorithm which is suited to a test scored by a neural network for a single criterion. In computerized adaptive testing, the item selection algorithm and the parameter estimation algorithm may be separated from the rest of the mechanics of testing. It is not necessary for these parts of the program to know about the content of the test, the specifications of the computer, or specific user

behaviors such as mouse movements. A fully operational program for adaptive testing must address these issues, but that is beyond the scope of this paper.

The algorithms here will be described in the language of mathematics. A subsequent section will outline the structure of a program to carry out these functions. Because of the ultimate goal of constructing an operational system, approximate solutions will be given in some cases to improve computational efficiency; although elegant solutions may be described, these approximations will be preferred.

Any item selection algorithm has three basic functions, each of which evaluates according to a rule. First, there must be a rule for selecting the first item, such as "Present item #1" or "Present the item with a difficulty closest to the mean ability level in the population." This may be a special case of, or separate from, the second rule, which governs how subsequent items are selected when some information is known about the examinee.

The third rule governs when to stop presenting items, and may be as simple as "Stop presenting items when ten items have been presented." Alternative stopping rules, however, may include a maximum standard error with which an examinee may leave the test. When the examinee is measured to that precision or better, the test ends. (Thissen & Mislevy, 2000) Some authors have argued in favor of fixed-length tests rather than fixed-precision tests, on the basis that an examinee who fails the test after a small number of items may feel that he has not been measured adequately to justify his failure, particularly in high-stakes contexts (see, e.g., Bergstrom & Lunz, 1999). This argument is not relevant to all testing circumstances, however. Either way, when the stopping rule

executes, the testing program must be able to produce a score (or at least a pass-fail judgment) and a measure of either reliability or error of measurement.

This chapter will primarily be concerned with the second rule, the continuing rule or next item selection algorithm. Before acceptable specific rules for a selection algorithm may be devised, however, it is necessary to consider the estimation procedure, which maintains the score and error estimates.

Specifically, it is necessary to observe the behavior of the estimates produced when some of the input data are held constant and others vary, representing the situation in which some values are uncertain. A series of increasingly complex examples will be presented to illustrate these behaviors.

In all of the examples that follow, a neural network is trained on a list of B biodata variables such as credentials and job experience ("biodata"), a list of I Likert-scaled or multiple choice items ("items") which may take on any of V integer values, and a list of S continuous-valued scales ("scales") with mean zero and standard deviation one. All adaptation will occur in the items and scales. The biodata questions will always be presented, as to do otherwise might be to ignore legal or functional requirements. To achieve maximum benefit from the adaptive process, the biodata questions will be presented first.

The neural network has a three-layer perceptron architecture; alternate architectures will require some re-derivation. Specific requirements of the neural network architecture will be noted.

2.1. Example 1: All data present but one item

In this particular case, all data is presented to the fully trained neural network except for one item, $i \in \{1, 2, ..., I\}$. Disregard for the moment how this one item was chosen to be omitted. Assume also that the biodata can be represented by a vector **B** of integers, and that the information resulting from the administration of S scales can be represented by an S-dimensional vector $\hat{\theta}$. That is, both are point estimates recorded with no uncertainty. Despite the estimation notation, $\hat{\theta}$ here is the final value, equivalent to the value on which the neural net was trained, and may as well be the true value because its uncertainty has been discarded.

The item may take on any of V values, leading to V different input patterns which may be presented to the neural network if the last item is presented. Each of these V input patterns will cause the neural net to produce an output; these outputs may be the same or different. Select one value of this item, v_i. Then v_i has a probability

$$p_{vi}=P(v_i|\hat{\theta}, \mathbf{B}, \mathbf{v}_{j\neq i}) \qquad (2)$$

where $\mathbf{v}_{j\neq i}$ is the vector of the I-1 known item responses. Given each complete input pattern, the neural network produces a value y. It follows that the distribution of predictions output by the neural network will have $Y \leq V$ possible values, because two input patterns may generate the same output pattern, but each input pattern results deterministically in a single output pattern. The probability of output y, drawn from this Y-valued set, will be

$$p_y = P(y) = p_{vi} * P(y|v_i, \hat{\theta}, \mathbf{B}, \mathbf{v}_{j \neq i}). \quad (3)$$

P(y|v) is, in this case, a binary value: is the output of the neural net equal to y given the specified input values, including v_i? The probability notation is used for consistency with subsequent examples.

Two descriptions of the output distribution are needed for either the next-item procedure or the stopping rule to evaluate. The first is a point estimate of a measure of central tendency, such as the mean value in continuous cases or the most likely value in discrete cases. When the stopping rule executes, this value will be returned as the score. An estimate of measurement precision is also needed; the next-item procedure to be developed will depend on changes in this quantity. The variance of the output distribution serves this function in continuous cases, and is mathematically convenient. In our example case, the mean corresponds to

$$\sum_y (y * p_y) \quad (4)$$

and the variance is

$$\sum_y ((y * p_y)^2) - (\sum_y (y * p_y))^2. \quad (5)$$

Although the mean given above is equal to the network's prediction of the criterion, the variance is not representative of the imprecision of that prediction. It is a measure of the uncertainty surrounding the examinee's final score if the examinee completed the entire assessment. This variance may be added to the variance of the

criterion expected for examinees whose final scores are equal to that mean value; the result is the expected variance of the criterion given the current best prediction.

2.2. Example 2: Two items missing

With the presentation of the last item thus modeled, consider the presentation of the second-last item from the pool. This item has *V* possible values v_h, and for each of these, the *V* values of the remaining item lead to several possible outputs as described above. Define *Y* now as the set of possible outputs resulting from the *V***V* possible response combinations to the two remaining items. We may still say that v_h has a probability

$$p_{vh}=P(v_h|\hat{\theta}, \mathbf{B}, \mathbf{v}_{j\neq h,i}). \quad (6)$$

Simillarly, each possible output still has probability

$$p_y=p_{vh}*P(v_i| v_h, \hat{\theta}, \mathbf{B}, \mathbf{v}_{j\neq h,i})*P(y| v_h, v_i, \hat{\theta}, \mathbf{B}, \mathbf{v}_{j\neq h,i}). \quad (7)$$

While this equation appears unfriendly, it may be simplified considerably if certain assumptions are met. Two cases are both likely and useful to consider.

In the first case, the *I* items which are not members of subscales are uncorrelated. This is the ideal case from the standpoint of the neural net; it means all redundancy has been accounted for by the use of the subscales. If the stand-alone item responses are statistically independent of each other and of the subscales. then $P(v_i| v_h, \hat{\theta}, \mathbf{B}, \mathbf{v}_{j\neq h,i})$ will be equal to $P(v_i|\mathbf{B})$; this distribution of responses will be constant regardless of how many

or how few other responses have been made. $P(v_i)$ could be independent of **B**, but this is not of great import as **B** is known prior to administration of the adaptive test.

In the second case, the *I* items are related to each other and to the scale scores only by a common factor, which may be a nuisance variable. (If the common factor is not a nuisance variable and the correlations are strong, CAT based on testlets and item response theory may be a better solution.) This is the case if, say, the items are susceptible to social desirability ("faking") effects. Examinees may be more or less inclined to present themselves favorably. This results in low but positive correlations between items in the socially desirable direction, even if those items are not all oriented the same direction in terms of the criterion. In this case, analytic computation of the outcome distribution is less straightforward, but still better than the general case.

2.3. Example 3: Many items missing

By induction, the formulae developed for one and two missing items may be extended to the case of an arbitrary set of items missing. Define $\mathbf{I_k}$ as the set of item responses known, and $\mathbf{I_u}$ as a set of responses that may be made to the remaining items. Then

$$p_y = P(y|\hat{\theta}, \mathbf{B}, \mathbf{I_k}) = \sum_{I_u} (P(y|\, \mathbf{I_u}, \hat{\theta}, \mathbf{B}, \mathbf{I_k}) * P(\mathbf{I_u}\,|\hat{\theta}, \mathbf{B}, \mathbf{I_k})). \qquad (8)$$

Analytic evaluation of the mean and variance of the expected outcome distribution becomes impractical quickly, particularly in the case where inputs may be correlated. A numeric approximation can be constructed with arbitrary precision.

The method of multiple imputations, attributed to Rubin (1987), was developed to handle missing data in statistical models. It calls for the substitution of "plausible values" in place of missing data, rather than a default value such as the mean of each distribution. Plausible values are random numbers which are scaled to the input ranges or recoded to the input values, and then filtered according to the input distribution. Computation based on this substitution is imputation; the "multiple" part of the method comes in when the computation is repeated with numerous sets of plausible values. Multiple imputations give an approximation of the expected outcome distribution.

In a procedural sense, the use of imputation operates as follows. Two random numbers, drawn from a uniform distribution between zero and one inclusive, are generated for each item which is missing. The first is converted into an admissible value for an item response. The second is compared without transformation to the expected probability of that item response. If it is lower, the value is accepted as plausible; if it is higher, it is discarded and new values are drawn.

The preceding description implies that each value is accepted or rejected separately. This is the case if and only if the remaining items are assumed to be independent of each other when conditioned on the known values. This is true if the items are actually independent, and approximately correct when the items are related only by a common factor. In the latter case, the expected distributions of each item are adjusted

based on the level of the common factor estimated from the observed data. Whether this adjustment is made based on item response theory, linear regression, or another technique is of little importance so long as this correction is small.

If the items are not conditionally independent of each other, plausible values must be accepted or rejected jointly. This is much more computationally intensive. Also, in this case, representing the joint probability distribution is complex and requires very large amounts of data; Zhou (1998) recommends using a neural net as the filter device, trained to predict the plausibility of sets of values.

Once an acceptable set of plausible values has been obtained, the observed and plausible values are fed to the neural net as inputs, and an output value is calculated. This procedure is repeated, each time with a new set of plausible values, for a specified number of iterations *N*. The result is a sample of *N* data points drawn from the distribution of output values which may be expected for this examinee. The mean and variance of this sample estimate the mean and variance of the theoretical distribution, and may be used in their place for the selection algorithm's calculations.

2.4. Error of measurement, and a candidate item selection algorithm

With these procedures, at any given time during the test, an estimate is available of the error of measurement, not from the true score or the actual employment outcome, but from the value which would be obtained if the entire test were administered. This error is expected to decrease monotonically as additional items are administered, and

becomes zero when the last item is completed. It is possible and useful to quantify this decrease.

Let item i be any item, but not the last available. Let $\mathbf{I_k}$ be the set of responses to items administered; $\mathbf{I_k}$ may be the null set. Let $\mathbf{I_u}$ be the responses that will be given if and when each additional item is administered, not including i. The incremental reduction in variance due to administering a shorter test when item i is administered is equal to

$$\begin{aligned}
&\mathrm{Var(current)}-\sum_i p_{vi}*\mathrm{Var(with\ v_i)}\\
&= \sum_y((y*P(y|\hat{\theta}, \mathbf{B}, \mathbf{I_k}))^2)-(\sum_y(y*P(y|\hat{\theta}, \mathbf{B}, \mathbf{I_k})))^2\\
&\qquad -\sum_i(p_{vi}*(\sum_y((y*P(y|\hat{\theta}, \mathbf{B}, \mathbf{I_k}, v_i))^2)-(\sum_y(y*P(y|\hat{\theta}, \mathbf{B}, \mathbf{I_k}, v_i)))^2)). \qquad (9)
\end{aligned}$$

Solving this equation requires estimation of V+1 variances by separate imputation. One is the current variance; the other V are estimates of what the variance will be if the examinee selects one available response.

On the basis of this model, a candidate rule for selecting subsequent items may be proposed. The rule may be stated as, "Choose the item which, in expectation, reduces the variance of the output by the greatest increment."

Computationally speaking, this requires a form of look-ahead procedure. For each remaining item, estimate the incremental reduction in variance, delta-variance, according to the formula already given. Choose the item with the highest delta-variance. Then discard the list; once another item is administered, the second-most-informative remaining item may not become the most useful. This situation does not require a violation of local independence to exist.

If there are I_u items remaining, the incremental reduction in variance must be estimated for each one. Although each incremental reduction calculation requires V+1 error variance estimations, the look-ahead procedure only requires I_u*V+1, because the current variance estimate may be re-used. Nevertheless, because each estimation by multiple imputation involves a large number (say, a thousand) neural network predictions, the procedure is computationally demanding. Nor is it amenable to pre-computation, because of the complex relationships that may exist between items and biodata. A look-up table for a five-item-long test from an item pool of thirty might easily have over twenty-four million cases, and that number scales exponentially with the length of the assessment.

2.5. Uncertainty in latent trait values

Thus far, the scales have been represented only as a point estimate, a vector of S exact values. No attention has been paid to how those values were calculated, or how many items have been asked from each scale. Because the scales are known to measure univariate constructs, it makes sense to estimate them using item response theory. One of the advantages of IRT-based estimation is the ability to report the error associated with such an estimate, or even a probability distribution for the location of the true latent trait value. Let us consider the latter possibility. For S scales, arbitrarily correlated, $\hat{\theta}$ is now replaced by an S-dimensional continuous probability distribution,

$$p_\theta(\mathbf{x})=P(\theta=\mathbf{x}), \qquad (10)$$

that is, the likelihood of the true trait values being **x**, conditioned on responses already made.

The distributed form of $\hat{\theta}$ carries through the calculations demonstrated previously. The output values y are now not a list of exact values that may be produced, but a genuinely continuous distribution of unknown form. The mean of y becomes

$$E(y)=\int_{-\infty}^{\infty}(y*p_y)\,dy \,/\, \int_{-\infty}^{\infty} y\,dy. \tag{11}$$

The variance is

$$Var(y)=E(y^2)-E(y)^2. \tag{12}$$

The sums over possible values of missing data must be integrated across all values of **x** before comparison, complicating the analytic form further. The difficulty of approximation by the method of multiple imputations is nearly unaffected, however. In a numeric approximation, an integral is just another sum, and this extension simply calls for the inclusion of the elements of $\hat{\theta}$ on the list of plausible values to be drawn.

Because the latent traits measured by the scales are arbitrarily correlated, the candidate plausible values **x** for each $\hat{\theta}$ vector should be drawn and filtered simultaneously, according to their joint probability distribution function $p_\theta(\mathbf{x})$. However, the joint probability distribution function may not be known, particularly if multidimensional IRT methods are not used to model the items. The misfit of the implied joint function that results from drawing plausible values independently should be evaluated on a case by case basis. Where correlations between scales are low or not well

known, the degree of misfit may be no greater than that which stems from the assumption of an incorrect distributional form.

Let us return to the general discussion. Incorporating uncertainty in scale values, as is implied by representing them as distributions, permits a wider range of values of y by spreading out the formerly discrete possibilities along a continuum. It is fair to assume that as the uncertainty in the trait estimate increases, the uncertainty in the output will also increase, or at least not decrease.

At any point during the administration of the items in a given scale, that distribution may be passed along to the neural net. (In practice, most if not all neural net programs cannot accept a distribution of values as an input, but the algebraic form allows it.) As more items have been presented, the distribution becomes narrower; the error of measurement of that trait becomes smaller. If some subset of the items in a scale are to be presented, regardless of the mechanism, it is worthwhile to consider the incremental effect of input uncertainty on output uncertainty.

For simplicity, first consider the case where all items have been administered. Recall that the change expected in the output per unit change in a given input is the sensitivity to that input, and that the sensitivity $\frac{\partial y}{\partial x_s}$ is calculated as the partial derivative of the output with respect to that input. The exact analytic form of $\frac{\partial y}{\partial x_s}$ varies according to the form of the neural network. For any neural network with one hidden layer, define a_j as the activation of a hidden node, w_j as the weight of the connection between hidden node j and the output, and w_{ij} as the weight of the connection between input node i and hidden node

j. Define $g(\mathbf{a})$ as the transfer function of the output node, and $f(\mathbf{x}, \mathbf{B}, \mathbf{I})$ as the transfer function of a hidden node. Then

$$\frac{\partial y}{\partial x_s} = \left(\frac{\partial y}{\partial a_j}\right)\left(\frac{\partial y_j}{\partial x_s}\right) = g'(\mathbf{a})\sum_j (w_j * w_{ij} * f'(\mathbf{x}, \mathbf{B}, \mathbf{I})). \tag{13}$$

It follows that the variance in the output which is attributable to uncertainty in the input is

$$\sigma_i^2 = \int_{\mathbf{x}} p_\theta(\mathbf{x}) * \left(\frac{\partial y}{\partial x_s}(\mathbf{x}, \mathbf{B}, \mathbf{I})\right)^2 * (x_s - E(x_s))^2 \, d\mathbf{x}. \tag{14}$$

The incremental effect of administering each remaining component item to any of the *S* scales may be compared by computing *V* hypothetical $p_\theta(\mathbf{x})$ distributions, passing them through this formula, and comparing the averaged results to the existing scale-attributable variance, in much the same way as the effect of administering a stand-alone item was calculated. However, this places a computational premium on having the scales. An approximation can ease the computational burden greatly, while still being unlikely to result in the choice to administer an uninformative item.

If the uncertainty in the scales is small relative to the variation in scale scores across the population, it may be assumed that the output as a function of $\mathbf{x}$ is closely approximated by a hyperplane in the vicinity of $E(\mathbf{x})$, where $p_\theta(\mathbf{x})$ is high. This will certainly be true after some items have been administered, and may be true initially due to information from the biodata. The explicit scale-attributable variance function may be simplified with some loss of information by substituting $E(\mathbf{x})$ into $\frac{\partial y}{\partial x_s}(\mathbf{x}, \mathbf{B}, \mathbf{I})$ instead of

integrating across plausible values. The resulting scalar value may be multiplied by the incremental reduction in scale variance for an estimate of scale-attributable variance.

A more complex case is more likely. This is the case in which some stand-alone items have not been administered, and yet the incremental effect of uncertainty of each scale score is still needed. Assuming either the independence or common-factor cases for item intercorrelations, the exact formula requires weighted summation across the possible values of $\mathbf{I_u}$ according to their conditional likelihood, as well as integration across $\mathbf{x}$.

The approximate formula may be estimated by the method of multiple imputations, or, because an estimate of uncertainty of this value is not required, a point estimate of $\mathbf{I_u}$ may be used. $E(\mathbf{I_u})$ may be an obvious candidate, following the use of $E(\mathbf{x})$. However, recall that the elements of $\mathbf{I_u}$ are responses to items which may be ordinal or even categorical. In either of those cases, the arithmetic mean may be an inadmissible value, or result in an output which is not actually "in the middle." The modal value of $\mathbf{I_u}$ is more appropriate. In both the independence and common-factor cases, this value may be easily obtained by taking the value of each element with the highest conditional probability.

2.6. A better item selection algorithm

The approximation of the effect of scale uncertainty on output uncertainty leads to a next-item selection rule, but it is not complete. It begins and ends at the level of the scale. That is, the selection algorithm accepts an estimate of reduction in scale variance

for each scale, and returns a decision about which scale, if any, to "spend" an item on. It does not control which item within the scale is administered, or consider how that reduction in variance may be achieved. Under this rule, a subordinate function must administer an item, return a posterior distribution as a component of **x**, estimate the reduction in scale variance from administering the next item (but not do so), and make a standing request for permission to actually administer that item.

If the posterior distribution is to be estimated using IRT from some form of unidimensional item model, it makes sense to use a traditional CAT to select the items within the scale. A CAT maintains a posterior distribution, which in the classic case is usually a list of values of p_θ associated with values of θ. It selects the next item based on a maximum posterior precision method, and estimates the variance of the posterior distribution after that item is administered based on a look-ahead procedure. The estimate can be carried out once, without reference to what happens between when it administers one item and the next, because a traditional, unidimensional CAT does not accept information from other scales. This is a feature, not a bug; it simplifies item modeling. Altogether, this estimation of scale variance reduction is computationally cheap.

The candidate rule for first and subsequent item selection may be revised into a cyclic procedure as follows: "For each scale, retrieve the expected reduction in variance from administering the next item, and multiply it by a point estimate of the sensitivity. For each stand-alone item, obtain the expected reduction in variance by simulating each possible outcome. Choose the item or scale which, in expectation, reduces the variance of the output by the greatest increment when one item is administered. If an item is chosen,

administer it and update $\mathbf{I_k}$. If a scale is chosen, the subordinate CAT should administer the pre-selected item, update **x**, select another item for maximum posterior precision, and 'try out' the next item to obtain the expected reduction in scale variance. The subordinate CAT should retain this value."

2.7. Modularity

Although this selection procedure has been developed in a relatively specific context, many features of the context may be changed without fundamentally altering the selection algorithm.

All of the mathematics above have been derived without reference to any specific mechanics of the neural net, other than example sensitivity functions. In fact, this procedure does not require that the predictive function be a neural net. Any mechanism will do if its output is a continuous, analytically differentiable function of the continuous inputs given any values of the discrete inputs. These are the functions well-modeled by neural nets, but no part or form of neural net calculations, nor any mechanism of fitting the model, is required for the technique to work. Note that some models can be considered special cases, which simplify the calculations -- sometimes to the point where the test is no longer adaptive. Multiple linear regression is one such model type.

The rationale for using subscales where items exhibit local dependence has been given, but subscales may simply be omitted if the item pool is appropriate. In some cases, testlets may be used instead of subscales, if the item content calls for it. These are

arbitrarily scored groups of locally dependent items which are always administered together. The selection rule for items can easily be adapted to penalize testlet-associated reduction of variance proportionally to the length of the testlet.

If subscales and/or testlets are used, stand-alone items may be omitted. This can easily occur in more theoretically well-defined areas of testing, such as academic assessment. This simplifies calculations considerably; the predictive relationship is essentially a guide to arbitrating between several univariate CATs competing for an examinee's time. In this case, however, building a fully multivariate CAT with joint estimation may be more effective.

Biodata, or rather, a pre-existing classification of the examinee which contributes information to item selection, is not necessary for this procedure. In applications other than an employee selection context, it may be considered more appropriate to use only population characteristics as a prior distribution. This decision has been made before in educational contexts (Mislevy et al, 1992)

3. Implementation and program structure

For a computer to administer a test, the structure of the test must first be programmed. This requires us to make explicit not only the mathematics of scoring but the functional operations of choosing and presenting items, recording and processing data. This chapter will discuss the structure of a program which administers an adaptive test. A prototype system was constructed according to this structure, and the results of its use will be shown in the next chapter.

First, let us consider the general architecture of such a program in terms of processes, as shown in Figure 2. The processes described are an extension of the three rules from Chapter 2: the starting rule, the continuing rule, and the stopping rule.

The starting rule is as follows: Begin a new log. Administer any fixed content, one item at a time, then go to the continuing rule.

Administering fixed content is, of course, its own trivial loop: Administer a biographical item. Is there another biographical item? If so, repeat. If not, go on. However, the structure of the fixed content administration may be much more complex than this without any effect on the final product; it is also not of particular interest.

The continuing rule is cyclic: Test for the stopping condition. If the stopping condition is satisfied, go to the stopping rule. Otherwise, select an item according to the item selection rule. Display the item, record a response and update the relevant internal structures. Estimate a score according to the scoring rule. Then, go to the continuing rule.

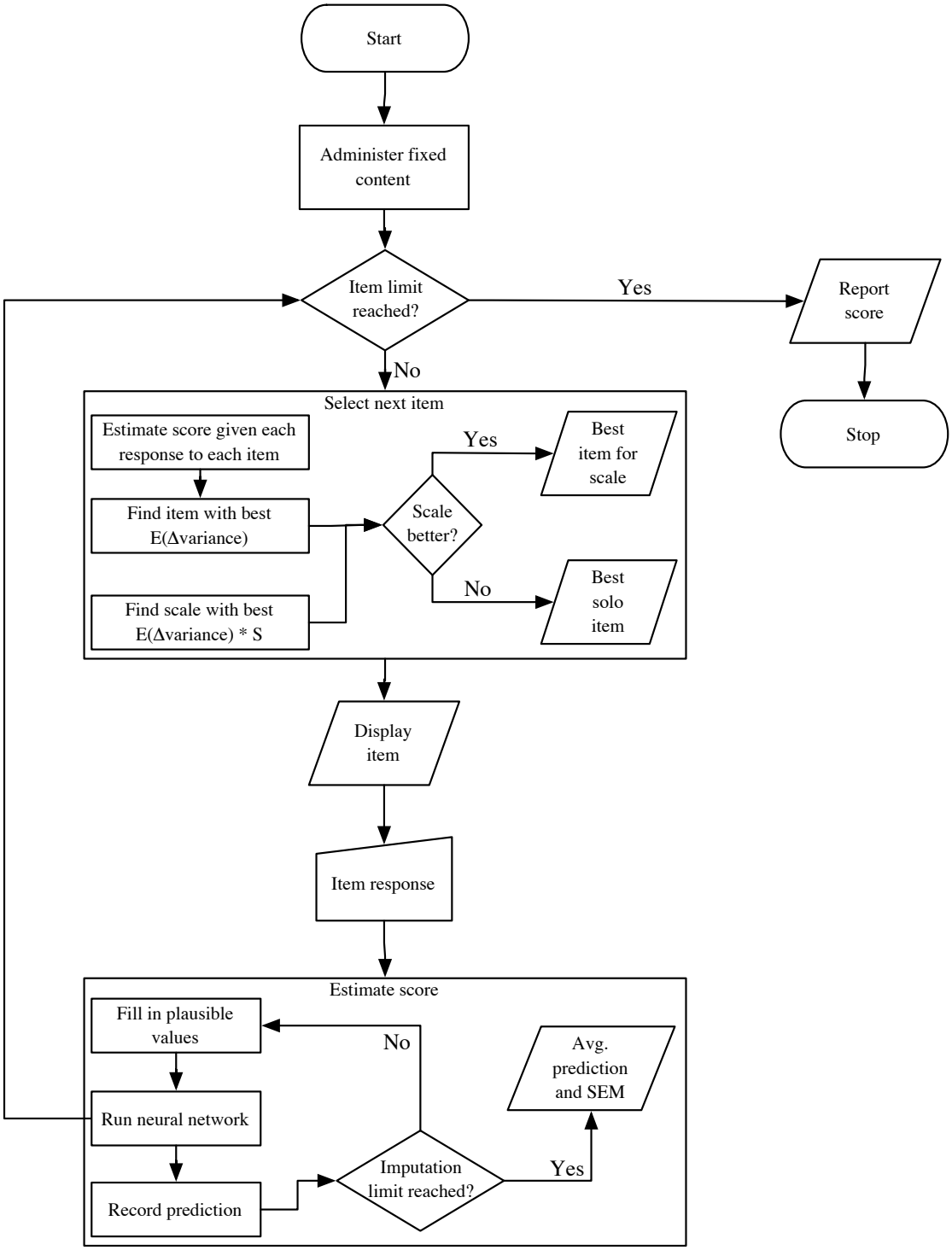

Figure 2. Architecture of the program, from a process standpoint.

Recall that the stopping condition may be the attainment of a specified precision or length of test, or another testable proposition. Regardless, when the condition is satisfied, the procedure for stopping is administrative. Report the score to the hiring manager. Thank the applicant. Save the log files.

The complexity of the CAT lies one level down, in the item selection rule and the scoring rule. Recall the item selection rule from Chapter 2. "For each scale, retrieve the expected reduction in variance from administering the next item, and multiply it by a point estimate of the sensitivity. For each stand-alone item, obtain the expected reduction in variance by simulating each possible outcome. Choose the item or scale which, in expectation, reduces the variance of the output by the greatest increment when one item is administered."

The scoring rule may be stated more simply. "Estimate the mean outcome if this applicant is hired, by feeding the neural net the known responses and different plausible values of the remaining data."

Another way to look at the architecture of such a program is to consider the flow of information between functional units, each of which maintains or accesses some data structures and performs specified functions. This view, as shown in Figure 3, displays more of the complexity inherent in CAT, and particularly in a hybrid CAT. Each box represents a functional unit, labelled with a name and then in some cases the primary data structure maintained by that functional unit. Each arrow represents the flow of mathematically important information. Requests and function calls are not shown.

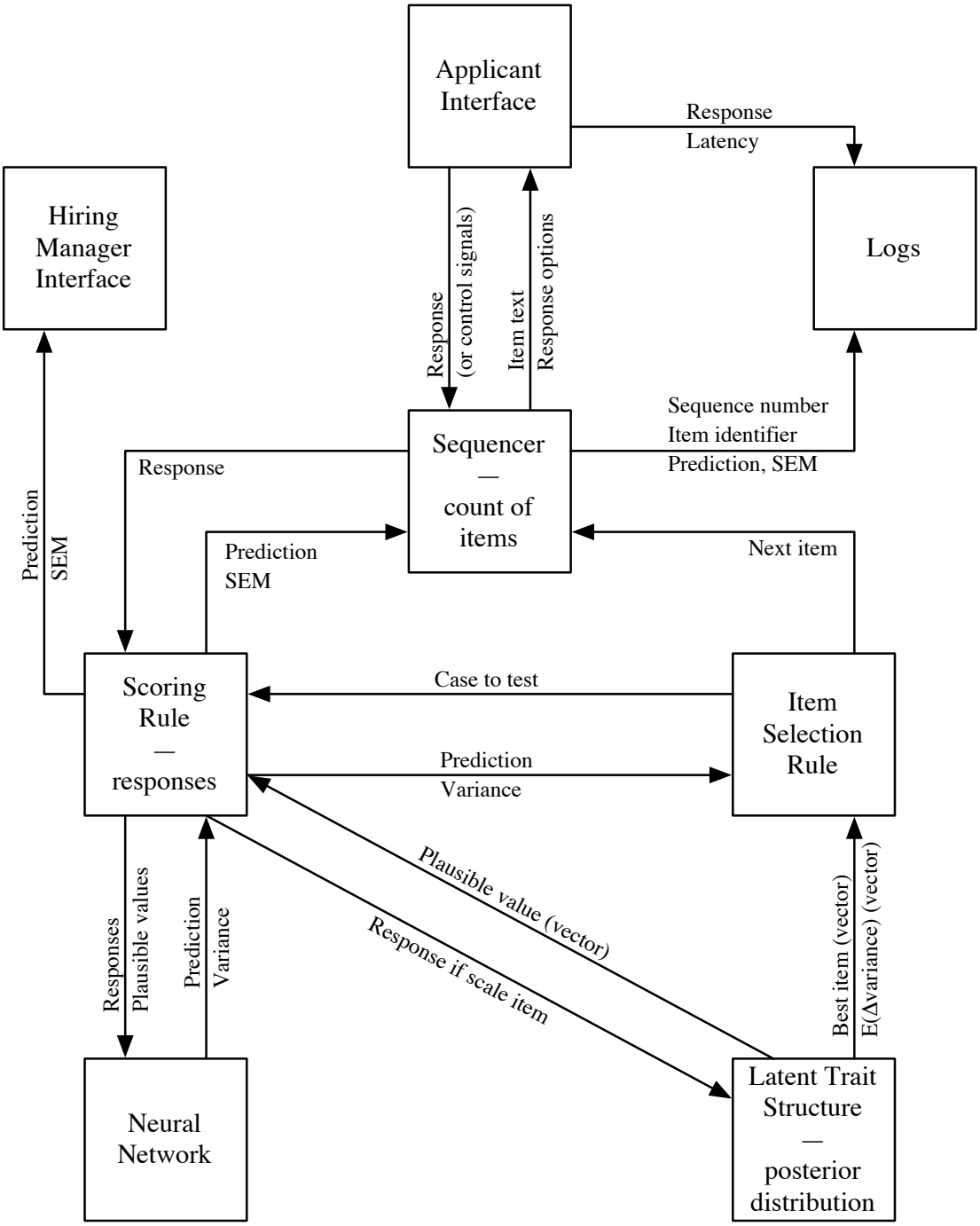

Figure 3. Architecture of the program, from a data flow perspective.

It is worth noting that the same structure, constructed for a traditional IRT-based CAT, has one fewer unit and four less connections. The neural network is not present and the item selection rule is not connected to the scoring rule. The remainder of Figure 3 is consistent with any computer-administered test, although in some cases where minor variation may be expected, choices have been made in order to be consistent with the simulation implementation. For example, the stopping rule may be a specified precision, a score may be reported to the applicant, or an error of measurement may not be available (as in the case of a fixed test).

Each of the functional units will be described in greater detail below.

3.1. Applicant Interface

The applicant interface has few, simple functions. It must allow the applicant to begin the test, which means telling the sequencer to initialize and the log keeper to open a new file for the applicant. The applicant should also be able to abort the test in an incomplete state if necessary. The interface is not obliged to be nice to the applicant in this case, but should at least reset itself for the next applicant.

Mostly, however, the applicant interface is there to present items, instructions, and information such as legal statements, and to allow the applicant to respond to open-ended as well as menu-type items. It must record the applicant's responses and response latencies to the logs, as well as passing the responses to the sequencer.

Beyond having enough screen space to display the whole item, the applicant interface is not subject to many restrictions. To avoid interfering with the measurement being attempted, it should be simple and clear. It may be necessary to prevent the applicant from multitasking, or requiring the computer to multitask. There are performance reasons for dedicated attention on both sides of the keyboard; performance issues will be discussed further at the end of this chapter.

3.2. Sequencer

The sequencer is responsible for deciding when to invoke the starting, stopping, and continuing rules, as well as organizing the events within the continuing rule. The sequencer keeps a running count of items, or keeps track of the error of measurement, as required by the stopping condition. It is also the primary source of data to be sent to the logs: the date and time started, the sequence number of the current item, the identifier and content of the item chosen, and the applicant's score.

When this CAT is implemented in a procedural language, the sequencer function calls and dismisses the item selection rule and scoring rule each time the continuing rule loops; it is thus unfortunately responsible for maintaining, disseminating and recovering a number of major data structures that it otherwise does not use, such as the posterior distribution vectors. It is more convenient for the purposes of discussion to associate those data structures with specific functional units at the "back end" of the program; I will borrow from object oriented programming to refer to different functions and persistent

data structures attached to agents such as the item selection rule. This is intellectual sleight of hand; I will also speak in procedural terms when it is more convenient to do so.

In the continuing rule loop, the sequencer tests the stopping condition. If the condition is not met, the sequencer asks the item selection rule for the next item and waits. Upon receiving an item identifier, it reports it to the log, tells the applicant interface to get a response, and waits. Upon receiving a response from the applicant interface, it passes the response to the scoring rule, asks the scoring rule for a score, and waits. Upon receiving a score, it passes the score on to the logs, then returns to the beginning of the loop.

3.3. Logs

The logs barely require their own functional unit at all. This agent is responsible for ensuring that all data passed to it is stored in an organized, safe and secure way. This may involve writing to a file, a database, or another structure.

The logs receive data including item identifiers, responses, latencies, and scores on an ongoing basis from the applicant interface and sequencer; in order to comply with possible court orders, the data must be recorded such that they cannot be lost even if the test is unceremoniously aborted, the power fails, or some other part of the program crashes. The current simulation does not achieve this goal, but an operational system must.

3.4. Item Selection Rule

The item selection rule is invoked by the sequencer. Its basic modus operandi is to acquire two pieces of information, make a comparison, and spit out the identifier of an item. It does not maintain any data structures of its own from iteration to iteration.

The two pieces of information it needs are the best possible expected reduction in variance due to administering an item, and the same quantity due to administering a scale. It does not matter which one it calculates first; they could be simultaneous if the language and supporting system permit threading. When both values are known, they are compared, and the item associated with the higher value is returned to the sequencer.

The best scale is chosen according to the method described in Chapter 2. In short, the item selection rule asks the latent trait structure for a list of the best items from each scale, and the expected reduction in scale variance associated with each one. Then it multiplies each by the sensitivity of the output to that input and finds the highest result. The sensitivity is not easy to calculate; for some parts it may be easier to run the neural network and record the final activations of the nodes.

The best item is also chosen as described in Chapter 2. This, however, requires trying out each possible response to each yet-unadministered item by submitting the current responses plus that one to the scoring rule. The variance for all responses to each remaining item are averaged, using weights corresponding to response probabilities, and subtracted from the current variance (also calculated by the scoring rule) to produce the

expected reduction in variance. The best item is the one associated with the highest expected reduction in variance.

3.5. Scoring Rule

The scoring rule may be invoked either by the sequencer or by the item selection rule. In these two cases, it behaves essentially the same, but for different purposes. In either case it provides a prediction and an error of measurement. The primary difference is that the sequencer needs the prediction made for the current state of known responses, while the item selection rule asks about a hypothetical set of responses. The sequencer is likely to want the error of measurement as a standard error, whereas the item selection rule uses a variance, but it is easy to alter either functional unit to reverse the transformation if the scoring rule is programmed to only give one type of response.

The scoring rule maintains a list of what response has been given to each item, and the current best prediction with error of measurement. When the sequencer reports a new response, the scoring rule determines whether it belongs to a stand-alone item or to a scale. If it belongs to a stand-alone item, the rule updates the list. If it belongs to a scale, it passes the response on to the latent trait structure.

In either case, the scoring rule updates its score. It searches the list for default values, which represent missing data. A specified number of times, it copies this list and fills in where data is missing, according to the rules of imputation: it generates random values and filters them according to their likelihood. For the scale values, it asks the

latent trait structure to generate plausible values according to the same rules. When the copy comprises a complete set of inputs, the scoring rule submits those inputs to the neural net and records its response. Once the specified number of responses is accumulated, it computes the mean and standard deviation (or variance), records them and reports them back to the sequencer.

The same procedure is carried out when the item selection rule offers a hypothetical next response, except that an additional temporary copy of the current responses table is generated. This way, the actual current values can be reset at the end, so that the hypothetical response is not mistaken for a real one.

The scoring rule, either on its own (every time) or through the sequencer (once) also supplies the final score to the hiring manager.

3.6. Latent Trait Structure

The latent trait structure, which generally corresponds to the subordinate CAT referred to in Chapter 2, responds to either the item selection rule or the scoring rule, providing them with two quite different pieces of information. The latent trait structure maintains the posterior distribution.

The item selection rule uses two vectors maintained by the latent trait structure, the list of the best next item for each scale and the list of expected reductions in variance upon administering one item from each scale. Because these vectors are maintained, they need not be calculated at the time they are required. In fact, it is more efficient to update

these vectors, as well as the posterior distribution, each time a scale item response is passed over from the scoring rule.

The scoring rule also requires, at a different time, a list of plausible values, one for each scale. Plausible values are constructed by the same generate-and-filter method used by the scoring rule, using the posterior distribution for that scale to determine the likelihood of a given generated value.

The first time the latent trait structure is invoked, before any items have been presented, it generates a prior distribution. This is a different name for the same matrix which will later be called the posterior distribution; it need not be kept separate. Assuming the joint distribution is not known and the scales are treated as independent, this distribution can be written as a matrix with S rows. Each contains Q values, representing the height of the marginal distribution at Q quadrature points centered around 0, such as every 0.1 from -3 to 3. The heights are generated according to either the empirical distribution observed for each pattern of biodata, or a theoretically reasonable distribution, such as the normal distribution with its parameters adjusted according to the biodata.

Subsequently, for each response given, the item characteristic curve corresponding to that response is convolved with the marginal distribution for the corresponding scale. For this to work, the item characteristic curve should be represented as a vector of likelihoods according to the same quadrature; the product of each member of the two vectors can then be taken. The result is the posterior distribution for that scale, and the distribution matrix is updated with the new values.

The best next item for a scale s may be chosen by finding the highest expected information gain. The expected information gain is approximated as the dot product of the sth row of the posterior distribution matrix and each item's information curve. Item information curves must, of course, be represented as a vector of heights corresponding to the same quadrature.

For each scale, the expected reduction in variance corresponding to that best item is calculated. This is done by finding the exact reduction in variance associated with each possible response, and computing a weighted average according to the likelihood of each response. This vector, along with the list of best items, is maintained until the item selection rule needs it.

3.7. Neural Network

The neural network is fairly standard, and an out of the box configuration can be used. It does not maintain any data structures, although it requires a network of weights and biases which it generated in its training period. It takes a standard list of inputs on which it has been trained, and returns one or more predictions, one for each outcome it was trained to predict. We have not discussed any cases in which there would be more than one prediction made in a single run of the neural network. The neural network is unaware of uncertainty and does not output an error estimate; all imputation and aggregation of multiple trials occurs in the scoring rule.

The neural network computation has three parts, of which the middle part is an iterative loop. First, it must preprocess the inputs, for example dividing a categorical variable into a series of binary variables, one representing each category. The network may also need to normalize continuous variables into a small range near zero; if this occurs, it must be reflected in the sensitivity calculation.

Once the inputs are preprocessed, the activations of the neural network nodes may be computed, one layer at a time. This can be accomplished in fairly few lines of code, being a systematic weighted summation. Finally, the program must read off the value of the output node and deliver it back to the scoring rule.

3.8. Optimization

Even with the speed of modern computers, this system has sufficient complexity and makes sufficient demands on raw processing power that issues of optimization must be considered. As in the early days of computerized adaptive testing, it is important to consider what constitutes an acceptable delay between items, as this limits the calculations that may be done at that time. However, the calculations which are necessary to make the test effective must be completed within that time. A compromise must be developed, weighing the need for processor-intensive procedures against the increase in computational demand associated with them. Some suggestions follow for improving performance.

How much of a delay is permissible? One second? Two? Recall that current tests with the same purpose are administered over the internet. Between items, there is already a delay associated with data transfer and web page rendering, which does not come as a surprise to the applicant. The length of this delay is not measured, and depends greatly on the actual internet connection available to the applicant. However, it is likely that an additional second, or even few seconds, of processing would be lost in this expected delay.

The choice of a computing language is relevant. Initial prototypes of this CAT were developed primarily in the R language, with readability in mind. The neural net was the exception. It was already in C and called from R; standard code was generated by the neural network module of Statistica 6 (StatSoft, 2003), and it was unnecessary to duplicate its function.

R is an interpreted language, and has a great deal of overhead associated with each calculation. For a system which requires a large number of simple calculations, R is not efficient. The result of using mostly R was a system that took up to an hour to select an item, under moderate requirements, on a mid-range 2004 computer. It was quickly decided that it was necessary to port at least the primary functions of the test to a compiled language. It was convenient to use C, because of the neural net already being programmed in that language.

The number of imputations required to achieve consistent estimates of the likely prediction and error of measurement is likely to vary according to the structure of the neural net. One that fits the data well, with a wider range of sensitivity values, will

require fewer iterations to achieve reliable results. The prototype system was reduced to 500 imputations per estimation without incident. It is possible that a lower number would have been acceptable.

Another approximation that may be made more coarse for the sake of efficiency is the vector representation of each posterior distribution, item characteristic curve, and item information curve. If relatively few items are available for each subscale, it is unlikely that any given latent trait will ever be known to the precision normally associated with CAT. If fine distinctions on the order of a tenth of a standard deviation will never be made because of the items available, there is no particular reason that the resolution of the discrete representation should be greater. Two tenths of a standard deviation may well be acceptable, if one's interest is only in separating those applicants who are high on the trait from those who are low on it. This speeds up every calculation involving the posterior distribution, of which there are many.

There are further optimizations that can streamline the calculation and approach the "few seconds" performance necessary. In an operating environment that allows threading, the maintenance processes of the latent trait structure, including updates to the posterior distribution and the look-ahead procedure that gives the next item and expected reduction in variance, may be shunted to a second thread. If a second processor is available, as is not uncommon in recent years, it may be used, and the complexity of the subordinate CAT need not be as limited.

4. A simulation study

The preceding sections have proposed a hybrid, neural net-based CAT. The current section describes a prototype system and an experiment to confirm that it has the expected benefits of an adaptive test. That is, the test should be shorter with little loss of validity; "little loss" will be defined in relation to a uniform or random reduction of the test. The test should report its own error of measurement accurately. Finally, the test should not administer the same items to all applicants.

In order to verify that the hybrid CAT meets these requirements, a prototype must be developed, including a fully trained neural net. A partial simulation procedure, in which data from applicants who took the test under non-adaptive conditions is requested one item at a time by the adaptive test, permits immediate comparison within an individual of the effect of different testing procedures.

Data from 3,989 employment applications were used for the partial simulation. All applicants in the sample were hired at the national retail chain to which these applications were submitted; no criterion data was available for applicants not hired, so their data could not be used.

Performance data were collected over one month. The entire sample population was employed during that one month period and had been employed for at least one month.

The performance dimension measured was sales productivity. The dollar amount of sales attributable to an employee is routinely tracked by the company and compared on

a monthly basis to a sales goal. For this study, that dollar amount of sales was divided by the number of hours worked to provide a sales-per-hour figure. Sales per hour were then normalized within equivalent groups defined by job class, in order to limit the "noise" introduced by environmental factors not related to individual personality characteristics.

Each store employs several sales associates, and one or more cashiers, stockers, and managers. Sales associates made up the bulk of the sample, but the other jobs were also represented. There is expected to be employee movement between jobs, so it is not practical to extensively distinguish between the requirements of one job and those of another when considering a candidate for employment.

Slightly more than half the sample (50.1%) reported being male; 4.6% omitted the question. No single race made up the majority of the sample; 39% reported being African-American, and 37% reported being Caucasian. 4.7% omitted the question, and other races made up the remainder.

4.1. Source tests

All applicants responded to the same form of the Unicru Sales test, a test designed to predict success in floor sales through several behaviors.

The test was administered in one of two modes. Single-purpose kiosks were available inside store locations; the custom devices in the kiosks are referred to as "screen phones" (Figure 4). Applicants with access to the Internet could also apply at a Web site, and take the test within their Web browsers. The display capabilities of a screen phone are

not as sophisticated as those of a Web browser, but the input device is better defined. These technical differences required separate implementations of the test, and resulted in different user experiences. In addition, the device used to submit an application implies one of two test-taking environments: the store to which the application is being submitted, and a user-chosen location which likely afforded more privacy and comfort. Application mode was retained in order to provide context to other data obtained.

As its name might imply, the Sales test was expected to predict job performance in a customer-facing, selling environment. Dollar value of sales is a reasonable criterion measure against which to measure the Sales test.

Each of the tests measures several traits, on the principle that multiple behaviors may lead to the same business outcomes. According to an unpublished technical report, the Sales test was designed to measure sociability, dominance, adaptability, optimism, and the applicants' own estimates of their on-the-job effort and practical intelligence. These traits are implicitly assumed to be compensatory, but in an arbitrary fashion; the test was only loosely balanced to have equal numbers of these items, and was refined according to empirical correlations. (Paajanen, personal communication, 7/15/2004)

Figure 4. A screen phone.

Of the 80 items on the test, 49 were sorted into 7 reliable subscales and validated across multiple data sets and multiple organizations. The data set at hand was not used in subscale development. The apparent central constructs of the subscales and the expected constructs on the tests matched fairly well, but not perfectly. Most significantly, the applicants' judgments of their own ability and effort were highly correlated; the applicants had a general level of self-efficacy which they expressed on all obviously-valenced items. Whether this characteristic amounts to the desire to "fake good" or merely self-esteem, it was not separable into one opinion about ability and one about effort.

Other constructs, such as sociability, dominance and adaptability, were clearly separable. Dominance, in fact, had to be split into separate scales for leadership ambitions and leadership-relevant traits, correlated about 0.4. Because of the several distinct scales, a one-factor model was not supported for the overall test.

Thirty-one items remained as unique items after scales were constructed. These items represented a combination of items thought to be complex and items that tapped underrepresented constructs.

Of the numerous available biographical data, seven items were chosen according to the following pragmatic criteria. The items were required to have a finite (and small) number of possible responses, such as those chosen from a list; free response items were not allowed. Items about membership in protected classes were not used. Items were also not used if they could be used to identify the region from which an application originated; it is not useful to know whether New England employees perform better than California

employees, because positions must be filled in all regions. Of the items that passed those three tests, the highest possible amount of criterion variance they could explain was determined by an information theoretic procedure (Chambless & Scarborough, 2002); a list was made of those which were informative either singly or jointly. Highly collinear items were dropped from the list. Finally, one item was added which had been observed to have higher-order effects in a previous sample: application mode. The result was a list of seven biographical items.

4.2. Neural Network

For the present study, the sample was divided into one training sample and two holdout samples by independent random assignment of each case. 2,950 applications were assigned to the training sample; 648 and 391 were assigned to the holdout conditions, for an approximate 75/15/10 split.

Item parameters were obtained for the scales to be used by the subordinate CAT. Data for this process were drawn from a non-overlapping sample of 97,563 applicants at a retailer expected to have a similar sales environment. It was anticipated that hires at one or both chains might differ on the scale constructs, but applicants were likely to be similar.

The nominal model was applied to each group of items expected to form an internally consistent univariate scale. The nominal model is an item response model which predicts the likelihood of each of several responses, usually multiple-choice, given

the level of a single latent trait (Thissen & Steinberg, 1986). Although the items were Likert scales, the nominal model provided a superior fit compared to constrained models such as the rating scale model and graded response model.

A three-layer perceptron was trained on the training sample, using 7 scales, 31 additional items, and 7 biographical data as inputs, and 12 hidden nodes. The number of hidden nodes is not known to be optimal, but is not unreasonable given the number of training cases. The network was fully connected; weights were established through one hundred iterations of backpropagation, with a momentum coefficient of 0.3, followed by refinement through conjugate gradient descent. To avoid overfitting, on each iteration, noise was added to the inputs. The noise was distributed normally with mean 0 and standard deviation 0.1. The first holdout sample was also used to test whether overfitting had occurred.

After 100 iterations of backpropagation and 21 of conjugate gradient descent, the network appeared to have found either a local or global minimum; the fit of the network to the data stopped improving noticeably. Overfit was not evident; the correlation with actual outcomes was 0.123 in the training sample and 0.121 in the first holdout sample, so the network was accepted. The fit of the network to the data was relatively poor for this application, indicated by the low correlation in both the training and first holdout samples. However, the fit was sufficient that the network weights were likely to be meaningful.

4.3. Method

The effectiveness of the item selection method was tested on the first hundred cases from the second holdout sample, selected sequentially by application date. Predictions of per-hour sales were made for these cases under five conditions. In the "all data" condition, each case was fed to the neural net with no missing data and its prediction recorded. In the two "adaptive" conditions, a mock user interface submitted the required biodata items to the CAT, which was then allowed to choose a specified number of items (10 or 20) according to its methodology. As each item was chosen, the mock user interface reported the actual response to the CAT; a prediction was made without the remaining items. Finally, in the two corresponding "random" conditions, an equal number of items were chosen at random and the rest considered missing. Estimation in the random conditions was performed by the method of multiple imputations, as in the adaptive conditions, but the informed item selection routines were disabled.

4.4. Results

To ensure that this testing process has the expected benefits of an adaptive test, it is necessary to ask four questions. First, is a prediction following adaptive selection more accurate than one made following the same number of items administered at random? Second, is the error of measurement reported by the test program reflective of the actual error in estimation of the final prediction? Third, is the test in fact adapting, or simply

recognizing that certain items are universally more informative than others? Finally, how many items must be administered before the adaptive test delivers a reasonable approximation of the prediction made with full information?

To the first question, it may be conclusively stated that the adaptive item selection algorithm results in an improvement over random item administration. The absolute value of the difference between predictions in the adaptive and all data conditions was less than that between predictions in the random and all data conditions (Table 1; $p=0.03$ for 10 items and $p=0.0002$ for 20). The reported standard error of measurement was lower in the adaptive case at ten items and at twenty items (Table 2; $p<0.00001$ in both cases). Correlation with predictions in the all data case was higher for the adaptive case at both test lengths (Table 3; $p<0.05$ in both cases).

Is the error of measurement reported by the test program reflective of the actual error in estimation of the final prediction? One would expect the absolute differences between the test's predictions and the fully informed predictions, divided by the reported standard error of measurement, to be distributed with standard deviation one. At both test lengths, they were distributed with standard deviation 1.12, indistinguishable from 1 at 100 cases. In the absence of contradictory evidence, we may assume that the standard errors of measurement reported by the program are reflective of actual precision. Oddly, the partially informed predictions were biased toward a lower performance than the fully informed predictions. This bias may stem from the use of a prior distribution based on the applicant population for latent trait estimation in the cases of persons already known to be selected as employees. Some selection had been done for better traits, which was not

Table 1. Mean absolute difference from the "all data" condition.

Test length	Adaptive condition	Random condition
10 items	0.097 (0.084)	0.116 (0.099)
20 items	0.086 (0.074)	0.115 (0.099)

Table 2. Mean standard error of measurement as reported by the test.

Test length	Adaptive condition	Random condition
10 items	0.108 (0.017)	0.131 (0.009)
20 items	0.092 (0.020)	0.129 (0.010)

Table 3. Correlation with "all data" condition.

Test length	Adaptive condition	Random condition
10	0.60 (0.08)	0.21 (0.10)
20	0.70 (0.07)	0.22 (0.10)

taken into account by the test. The bias was lower in the 20-item case than the 10-item case, indicating slow convergence.

Is the test in fact adapting to individuals? It is possible for an item selection algorithm to outperform random item administration simply because some items are always more useful than others. In order to determine whether this is the case, one must examine the frequency of administration of different items. Only one item was given to

every applicant at both test lengths, and not always in the same ordinal position. Some items appeared relatively frequently, while 21 items never appeared in either condition, suggesting that there are some items which are more useful for a broad range of applicants than other items. This result suggests that the test is indeed adapting.

How many items are enough? In a practical situation, a decision must be made about how long the new adaptive test must be in order to deliver a reasonable approximation of the fully informed result. This decision hinges on what it means to be a reasonable approximation. The approximation will necessarily lower the criterion validity coefficient of the test, but is a reduction of 0.01 acceptable? 0.02? 0.05?

Let us assume that the true validity of the test is known to a certain precision, based on testing with a holdout sample. Let us then propose a rule of thumb: a reduction in validity which is less than the standard error of estimation of the validity coefficient is a reasonable approximation. By this rule of thumb, if the fully informed prediction had a validity coefficient of 0.20 with an error of estimation of 0.02, an adaptive test's prediction must correlate at least 0.90 with the fully informed prediction in order to be sufficient. If the neural net were trained to a validity of 0.30 with the same error of estimation, the prediction must correlate 0.93 in order to be acceptable.

In the demonstration case, the neural network was trained to a much lower validity, 0.12, atypical in practice. By the rule of thumb, the correlation of 0.70 achieved in the twenty-item condition was insufficient even at this level of validity. A longer test, for example 30 items, might be needed.

5. Conclusion

Over the course of the last three chapters, the mathematics, the architecture, and the performance of a computerized adaptive test have been demonstrated. In this dissertation, I have initiated the development of computerized adaptive tests that are based on underlying neural networks. Although I have focused on one specific implementation of this approach and its application to the problem of predicting sales performance, the methods here are easily generalized to other problems and alternative network architectures.

From another perspective, a neural network designed to recognize patterns leading to positive employment outcomes has been combined with a process that gathers the best possible information for improving its prediction, given constraints on the quantity of inputs allowable. The resulting hybrid is functioning according to the expectations placed on neural networks as well as those placed on adaptive tests.

It is highly empirical: it can model an arbitrary output function over an arbitrarily multidimensional input space. It is efficient: it achieves a much shorter test with relatively little loss of precision. It can report its own error of measurement: the error of estimation of a prediction can be scaled according to the validity of the prediction to give an error of estimation of the outcome. It permits comparison of applicants who did not answer the same items: it places them all on a common scale in terms of the predicted outcome, even if available item content is changed or the neural model is revised.

The neural network-based testing architecture devised here represents a first step into the domain of adaptive testing where multiple traits are simultaneously estimated. The prototype described in this paper maintains a latent trait structure involving seven separate traits, although it does not report a profile of scores. It would only be appropriate to report such a profile if each trait were thoroughly tested to attain a high test-retest reliability; while that did not occur in the example demonstrated and was not needed for the purpose described, it is a possible future case.

This hybrid, particularly the prototype implementation, is not elegant. It was developed for a practical purpose, and its success will be determined upon its ability to be used for that practical purpose. Much work must be done before an operational system can be deployed to real stores and real managers. This prototype holds promise; only time will tell if it is fulfilled.

References

Autor, D. & Scarborough, D. (2004) Will job testing harm minority workers? Working paper 04-29, Department of Economics, Massachusetts Institute of Technology. http://papers.ssrn.com/so13/papers.cfm?abstract_id=580941.

Barrick, M.R. & Mount, M.K. (1991) The Big Five personality dimensions and job performance: a meta-analysis. *Personnel Psychology, 44:1,* 1-26.

Barrick, M.R. & Ryan, A.M. (2003) *Personality and work*. San Francisco, CA: Jossey-Bass.

Bergstrom, B.A. & Lunz, M.E. (1999) CAT for Certification and Licensure. In Drasgow, F. & Olson-Buchanan, J.B. (Eds.) *Innovations in Computerized Assessment* (pp. 67-92). Mahwah, NJ: Lawrence Erlbaum Associates.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M & Novick, M.R. (ed) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 397-479.

Bishop, C.M. (1995) *Neural networks for pattern recognition*. New York, NY: Oxford University Press.

Bock, R.D. & Mislevy, R.J. (1982) Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6:4,* 431-444.

Braunfeld, P.G. (1964) Problems and prospectus of teaching with a computer. *Journal of Educational Psychology, 55(4),* 201-211.

Carroll, J.R. (1993) *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Chambless, B. & Scarborough, D. (2002) Information theoretic feature selection for neural network training. *Proceedings of the International Joint Conference on Neural Networks*. Washington, DC.

Collins, J.M. & Clark, M.R. (1993) An application of the theory of neural computation to the prediction of workplace behavior: an illustration and assessment of network analysis. *Personnel Psychology, 46:3,* 503-524.

Crick, F. (1989) The recent excitement about neural networks. *Nature, 337:2,* 129-132.

Dempsey, J.R.; Folchi, J.S. & Sands, W.A. (1995) *Comparison of alternative types of prediction models for personnel attrition*. Alexandria, VA: Human Resources Research Organization.

Drasgow, F. & Olson-Buchanan, J.B. (1999) *Innovations in Computerized Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S.E. & Reise, S.P. (2000) *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fraley, C. & Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97,* 611-631.

Garson, G.D. (1998) *Neural networks: An introductory guide for social scientists*. London: Sage Publications

Gelman, A.; Carlin, J.B.; Stern, H.S. & Rubin, D.B. (2004) *Bayesian data analysis (second edition)*. Boca Raton, FL: Chapman & Hall/CRC.

Guion, R.M. (1998) *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hagan, M.T.; Demuth, H.B.; Beale, M. (1996) *Neural network design*. Boston, MA: PWS Publishing.

Hartigan, J.A. (1975) *Clustering Algorithms*. New York, NY: John Wiley & Sons.

Haykin, S. (1999) *Neural networks: a comprehensive foundation*. Upper Saddle River, NJ: Prentice-Hall.

Hertz, J.; Krogh, A. & Palmer, R. G. (1991) *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley.

Hunt, E. (1995) *Will we be smart enough? A cognitive analysis of the coming workforce*. New York, NY: Russell Sage Foundation.

Hunt, E. (2002) *Thoughts on Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hunt, S.T. (1996) Generic work behavior: an investigation into the dimensions of entry-level, hourly job performance. Personnel Psychology, 49:1, 51-83.

Hunt, S.T. & Paajanen, G. (2003) The value of uncommon variance: designing personality selection measures for multi-dimensional predictor and criterion spaces.

Paper presented at the nineteenth annual conference of the Society for Industrial and Organizational Psychology, April 12, 2003.

Lord, F.M & Novick, M.R. (1968) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F.M. (1971) The self-scoring flexilevel test. *Journal of Educational Measurement, 8:3,* 147-151.

Matthews, G. & Deary, I.J. (1998) *Personality traits*. Cambridge: Cambridge University Press.

McDonald, R.P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Minsky, M.L. & Papert, S.A. (1969) *Perceptrons*. Cambridge, MA: MIT Press.

Mislevy, R.J. & Bock, R.D. (1989) A hierarchical item-response model for educational testing. In Bock, R.D. (ed.) *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Mislevy, R.J.; Johnson, E.G. & Muraki, E. (1992) Scaling procedures in NAEP. *Journal of Educational Statistics, 17:2,* 131-154.

Montaño, J.J. & Palmer, A. (2003) Numeric sensitivity analysis applied to feedforward neural networks. *Neural Computation and Applications, 12,* 119-125.

Muraki, E. & Carlson, J.E. (1995) Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19:1,* 73-90.

Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996) *Applied linear statistical models (fourth edition)*. Boston, MA: McGraw Hill.

Ployhart, R.E.; Lim, B.-C. & Chan, K.-Y. (2001) Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology, 54:4,* 809-843.

Reise, S.P. & Henson, J.M. (2000) Computerization and adaptive administration of the NEO PI-R. *Assessment, 7:4,* 347-364.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Scarborough, D. (2005) *Neural networks in organizational research*. Washington, DC: American Psychological Association

Schmidt McCollam, K.M. (1998) Latent trait and latent class models. In Marcoulides, G.A. (Ed.) *Modern methods for business research* (pp. 23-46). Mahwah, NJ: Lawrence Erlbaum Associates.

Schmidt, F.L. & Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124:2,* 262-274.

Segall, D.O. & Moreno, K.E. (1999) Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In Drasgow, F. & Olson-Buchanan, J.B. (Eds.) *Innovations in Computerized Assessment* (pp. 35-65). Mahwah, NJ: Lawrence Erlbaum Associates.

Somers, M.J. (1999) Application of two neural network paradigms to the study of voluntary employee turnover. *Journal of Applied Psychology, 84:2,* 177-185.

Somers, M.J. (2000) Self-organizing maps and commitment profiles. *Paper presented at the fifteenth annual conference of the Society for Industrial and Organizational Psychology.*

Somers, M.J. (2001) Thinking differently: assessing nonlinearities in the relationship between work attitudes and job performance using a Bayesian neural network. *Journal of Occupational and Organizational Psychology,* 74:1, 47-61.

Steinberg, L.; Thissen, D. & Wainer, H. (2000) Validity. In Wainer, H. (Ed.) *Computerized Adaptive Testing: a primer (second edition)* (pp. 185-228). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. (2000) Reliability and measurement precision. In Wainer, H. (Ed.) *Computerized Adaptive Testing: a primer (second edition)* (pp. 159-183). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Mislevy, R.J. (2000) Testing algorithms. In Wainer, H. (Ed.) *Computerized Adaptive Testing: a primer (second edition)* (pp. 101-132). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Steinberg, L. (1986) A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen, D. & Wainer, H.W. (2001) *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Turner, R.G. (1978) Consistency, self-consciousness, and the predictive validity of typical and maximal performance measures. *Journal of Research in Psychology, 12:1,* 117-132.

Vispoel, W.P. (1999) Creating computerized adaptive tests of music aptitude: Problems, solutions, and future directions. In Drasgow, F. & Olson-Buchanan, J.B. (Eds.) *Innovations in Computerized Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2000) *Computerized Adaptive Testing: a primer (second edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R.J. (2000) Item response theory, item calibration, and proficiency estimation. In Wainer, H. (Ed.) *Computerized Adaptive Testing: a primer (second edition)* (pp. 61-99). Mahwah, NJ: Lawrence Erlbaum Associates.

Wanek, J.E.; Sackett, P.R. & Ones, D.S. (1998) Item-level analysis of integrity: A judgmental approach to defining sub-factors. *Paper presented at the thirteenth annual conference of the Society for Industrial and Organization Psychology.*

Ware, J.E.; Kosinski, M.; Bjorner, J.B.; Bayliss, M.S.; Batenhorst, A.; Dahlof, C.G.H.; Tepper, S. & Dowson, A. (2003) Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12:8,* 935-952.

Wiberg, M. (2003) An optimal design approach to criterion-referenced computerized testing. *Journal of Educational and Behavioral Statistics, 28:2,* 97-110.

Zhou, Y. (1998) Neural network learning from incomplete data. TR1998-13, Department of Computer and Information Science, University of Mississippi. http://www.cs.wustl.edu/~zy/learn.pdf

Zickar, M.J.; Overton, R.C.; Taylor, L.R. & Harms, H.J. (1999) The development of a computerized selection system for computer programmers in a financial services company. In Drasgow, F. & Olson-Buchanan, J.B. (Eds.) *Innovations in Computerized Assessment* (pp. 7-34). Mahwah, NJ: Lawrence Erlbaum Associates.

VITA

Anne Thissen-Roe earned a Bachelor of Science degree in Psychology at the University of Illinois at Urbana-Champaign in 2001. In 2005 she earned a Doctor of Philosophy in Cognitive Psychology at the University of Washington. Currently she works as an industrial psychologist.