Available online at www.sciencedirect.com
ScienceDirect
Journal of High Technology Management Research 17 (2007) 125–137
THE JOURNAL OF HIGH TECHNOLOGY MANAGEMENT RESEARCH

Can experts really assess future technology success? A neural network and Bayesian analysis of early stage technology proposals ☆

Craig S. Galbraith [a,*], Alex F. DeNoble [b,1], Sanford B. Ehrlich [c,2], Doug M. Kline [d,3]

[a] University of North Carolina, Department of Management and Marketing, Cameron School of Business, 601 South College Road, Wilmington, NC 28403, United States
[b] San Diego State University, Department of Management, College of Business Administration, 5500 Campanile Drive, San Diego, CA 92182- 8236, United States
[c] San Diego State University, Entrepreneurial Management Center, 5250 Campanile Drive, San Diego, CA 92182, United States
[d] University of North Carolina, Department of Information Systems and Operations Management, Cameron School of Business, 601 South College Road, Wilmington, NC 28403-3297, United States

Abstract

This paper compares experts' assessments to a set of structural variables to determine whether each effectively predicts technology success. Using 69 homeland security and defense-related technologies, expert reviewers scored each technology on various dimensions as part of a government grant funding process. These technologies were tracked over 3 years and degrees of success recorded. Different predictive models were estimated using an artificial neural network technique, the Bayesian Data Reduction Algorithm, and two regression equations. The results suggest that experts provide little predictive power, and that a reasonable technology success model can be estimated using a limited set of structural variables.

Keywords: Forecasting; Technology commercialization; Expert assessment

1. Introduction

Interest in understanding the factors that lead to successful technology development and commercialization has been increasing exponentially during the last decade. Since Cooper's (1979) and Maidique and Zirger's (1984) seminal research, a large body of empirical "technology success" studies has developed that attempts to capture the critical factors that are positively correlated with successful new product innovation and commercialization (see Balachandra

☆ This study was supported by the Center for Commercialization of Advanced Technology (CCAT), a SPAWAR/Office of Naval Research supported technology transfer consortium located in San Diego.
* Corresponding author. Tel.: +1 910 962 3775; fax: +1 910 962 3815.
E-mail addresses: galbraithc@uncwil.edu (C.S. Galbraith), adenoble@mail.sdsu.edu (A.F. DeNoble), sehrlich@projects.sdsu.edu (S.B. Ehrlich), klined@uncw.edu (D.M. Kline).
[1] Tel.: +1 619 594 4890; fax: +1 619 594 8879.
[2] Tel.: +1 619 594 3956; fax: +1 619 594 8879.
[3] Tel.: +1 910 962 7552; fax: +1 910 962 3815.

1047-8310/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.hitech.2006.11.002

& Friar, 1997; Belliveau, Griffin, & Somermeyer, 2002; Cooper, 2001; Griffin, 1997; Montoya-Weiss & Calantone, 1994 for summaries).

But while there appears to be little empirical agreement regarding the specific success variables, there does, however, seem to be some consensus regarding the classes, or groupings, of these key success factors. Astebro (2004), for example, suggests a general consensus around the dimensions of market, technology, environment, and organizational characteristics. Similarly, using a factor analytical procedure on 54 different variables, Heslop, McGregor, and Griffith (2001) grouped the key commercialization success variables into the four dimensions of market readiness, technology readiness, commercial readiness, and management readiness.

Unfortunately, the vast majority of empirical efforts to examine technology development success in an *ex-post* manner also suffer from various bias problems. These include hindsight or retrospective bias where successful innovators and entrepreneurs will create a contemporaneous impression of what led to a successful technology development, rather than remembering the actual sequence of confusing and non-obvious events (Astebro, 2004; Christensen-Szalanski & Fobian, 1991). Other common problems of *ex-post* studies are survivor selection bias where only successful technology transfers are investigated or discussed (Elton, Martin, & Blake, 1996; Shane & Stuart, 2002), and memory decay where the technology developer simply can't remember the important decision points from the past (Carland, Hoy, Boulton, & Carland, 1984; Davidsson & Honig 2003).

Inevitably these biases result in fairly obvious factors, such as "having a superior product", "being aware of market demand", or "leveraging skills from a firm's core competence" being associated with successful technology development (e.g., Cooper, 2001). As such, this line of research has done little in providing early, *ex-ante* predictive models of future technological success.

But while much of the *ex-post* technology commercialization research is somewhat tautological in nature (i.e., superior products have greater market success), this research has provided the foundation for developing various scoring systems for early technology review. For example, in most structured approaches to technology development, such as the well-known Stage-Gate model (Cooper, 1993, 2001), the PACE process (McGrath, 1996), and a number of Stage-Gate variations that have a more technological orientation (Ajamian & Koen, 2002; Eldred & McGrath, 1997; Eldred & Shapiro, 1996), there is typically an early stage technology review process where a proposed technology is formally evaluated on criteria such as market attractiveness, future competitiveness, and technical merit. Similar formal reviews are made by equity investors when reviewing technology-based business plans and by granting agencies when reviewing Small Business Innovation Research (SBIR) and other R&D proposals (Ajamian & Koen, 2002; Cooper, 1998; Linton, Walsh, & Morabito, 2002; Ozer, 1999).

Many of these early stage assessments now involve some type of multidimensional scoring sheet or rating process (Cooper, Edgett, & Klienschmidt, 2002). In fact, within the past decade there has been a stream of complex technology readiness check-lists or calibrated scoring models designed for early stage, or "fuzzy" front-end assessments (Koen et al., 2002; Mock, Kenkeremath, & Janis, 1993). These include NewProd (Cooper, 1993) which incorporates a 30 item check-list, ProGrid (McCollough, 1998) which suggests nine dimensions of commercial readiness within a 3×3 grid, and the 54 item, 4 dimensional "cloverleaf" scale (Heslop et al., 2001). In addition, NASA has developed a series of technology commercialization assessment tools, such as those from the "Re-Invention Initiative Between Industry and Technology" (RIBIT) and "Long Evaluation and Analysis for Partnership" (LEAP) programs (Schoenborn & Fredette, 1997). The Japanese Patent Office (JPO) has the Patent Evaluation Index, while many large firms have developed their own, specialized early-stage assessment formats for internal use (Koen et al., 2002).

An underlying assumption of all technology assessment check-lists, however, is that individuals can accurately assess these characteristics in an *ex-ante* manner when evaluating an early-stage technology. But while tremendous effort has been placed on these early-stage technology assessments, little is actually known about the predictive capability of these scales or whether the individuals that actually evaluate early-stage technologies, such as venture capitalists or grant reviewers, have the insight, foresight, or skills to accurately assess the "technical merit" or "commercial potential" of such technologies.

Given these fundamental questions, and the hindsight and memory decay biases inherent in *ex-post* analysis, a few researchers are starting to examine technology commercialization or success within an *ex-ante* framework. Most of this research has examined the decision making process among equity investors, such as venture capitalists (e.g., Baum & Silverman, 2004; Shepard & Zacharakis, 2002; Zacharakis & Shepard, 2001). For example, Zacharakis and Meyers (2000) conducted a controlled experiment by providing a sample of experienced venture capitalists with different levels of information on 25 non-associated investments, and found that venture capitalists had at, best, a

prediction accuracy of less than 40%, and that this prediction accuracy actually decreased as more information was provided.

Other research is starting to expand this analysis beyond venture capital investments. For example, Astebro (2004) studied 561 R&D "low-technology" Canadian projects' evaluations from the mid-1990s, and tracked their success over time. The research suggests that the factors of technology opportunity, developmental risk, expected profitability, and intellectual property protection provided the greatest explanatory power.

The study presented in this paper follows the call for more research that examines technology assessments in a contemporaneous manner, thus minimizing the various biases associated with *ex-post* technology success studies. The present study empirically examines the ability of expert reviewers to assess the future success of various technology proposals. We examine actual "investment decisions", then longitudinally track the success or failure of these technologies over several years.

In addition, this research has a second objective. Because it is recognized that technology and business assessments are, at best, "fuzzy" in nature, it is likely that traditional linear models may not capture the actual predictive power of expert assessments. Therefore, we utilize two different techniques that are far less restrictive in their assumptions — a neural network technique and a Bayesian data reduction technique. The results of these two techniques are then compared with the results of two different traditional linear techniques, a step-wise regression analysis and a logistic regression model.

2. Methods

This research study examined 69 technologies funded by the Center for Commercialization of Advanced Technology (CCAT). CCAT is a congressional-funded Department of Defense (SPAWAR) technology transfer agency that has the specific mission of funding technologies being developed by small private sector R&D firms, government research agencies and university laboratories that have a specific homeland security and/or an anti-terrorism application. In particular, proposals for funding are received and reviewed approximately three times per year. Between 2001 and 2003 approximately 600 proposals were received and reviewed by CCAT, and 69 were funded; thus, the present study only tracks the funded technologies. Around 70% of both the received and funded proposals are from small firms, with the remaining split between university and government laboratories.

CCAT funding is considered supplemental (averaging around $75k) to other funding sources, and many of the applicants have previously received SBIR and other DoD funding. The technologies are generally in early to middle stages of technology development, ranging between proof-of-concept to lab prototype, with the CCAT funding specifically targeted to advancing the technology to the next phase of development. The technologies are classified as biometrics, communications, computers, electronics, life sciences, materials, photonics, and sensors, and can be considered highly advanced post 9-11 technologies. Examples of the technologies studied are "real time beryllium air particulate monitor", "new adjuvant class for enhancement of vaccinations", "specific multiplex diagnostic test for West Nile Virus", "cross-language information extraction", and "advanced gyroscopic infrared guidance systems". All of the technologies are considered to have dual-use capabilities for both government (e.g., military, homeland security) and commercial markets. Each received proposal is independently evaluated by a panel of experts, and scored, based upon the expert's assessments of reasonableness of request, technical merit, commercial potential, ability to sustain competitive advantage, and ability of the project team to execute the plan.

2.1. Variables

As a longitudinal study, we attempt to determine the predictive power of expert assessments obtained at the time of the proposal submission, and how they relate to various objective measurements of technology success obtained one and a half to three years later. The "expert assessment variables" reflect the individual expert's assessments of both the technology, and the technology developer organization, on five dimensions. In addition, several control variables were included in the analysis.

2.1.1. Expert assessment variables

Independent expert reviewers formally evaluated each received proposal on a number of dimensions defined by the CCAT review process. We used five variables: reasonableness of request (EA1), technical merit (EA2),

commercial potential (EA3), ability to sustain competitive advantage (EA4), and ability of the project team to execute plan (EA5), which directly correspond to the key dimensions reported by Heslop et al., (2001) and Astebro (2004). Reviewers used an 11-point Likert scale to score each variable. Scoring was done individually and independent of each other.

All reviewers, regardless of professional background, had substantial experience with technology-based ventures. Not only did each reviewer report their expertise in a particular class of technologies (such as biological warfare agents or biometric or materials engineering) prior to an application being sent to them, but after receiving the application the reviewers were asked to exclude themselves from the evaluation of any technology they did not feel qualified to evaluate. All the reviewers held either masters or doctorate degrees and included venture capitalists, scientists, engineers, technology transfer officers, technology commercialization consultants, and both private and government R&D managers. Each technology was evaluated by about three reviewers. Approximately 70 different reviewers were used over the time period studied, resulting in approximately 180 total evaluations of the technologies under investigation.

2.1.2. Control variables

The organizational control variables were obtained from information provided in the proposal packet. These variables were coded as categorical variables and included: age of firm (AGE: 0 to 3 years, 3 to 5 years, over 5 years), number of employees (EMP: 1 to 5; 6 to 20; 20 to 50; over 50), revenues (REV: none, less than $100k, $100k to $500k, $500k to $1m, over $1m), diversification level of firm (DIVER: single sector, related sectors, unrelated sectors), education of team (ED: highest level obtained by at least one team member, BA, Masters, Doctorate), stage of technology development (STAGE: concept, proof of concept, patent, lab prototype, field prototype, beta test, commercial), amount of external funding, (FUND: under $100k, $100k to $750k, over $750k), and formal strategic alliance or partner (PART: no, yes). For the technologies developed in government or university laboratories, we coded AGE as the age of the technology within the lab. For size, we assumed that the lab had access to the advantages of large organizational size and overhead support (for example, a technology transfer office, attorneys, staff support, etc.); thus, EMP was coded as "over 50". For DIVER the various technologies being developed by the research team were analyzed, and in most, but not all, cases were recorded as a single sector activity. Since there was a somewhat subjective coding assessment for the government or university laboratory sub-sample, our research estimated models for both the full sample and the sub-sample of just private firms.

2.1.3. Success variables

For the purposes of this study, the success of the technology was evaluated in late 2004, or approximately one and a half to three years after the initial evaluation. The success metric (SUCCESS) was an objective, market based measure of success, that is, whether or not a cooperative research and development agreement (CRADA) or license was signed, the technology received additional SBIR or equity funding, or an actual product was launched. SUCCESS was coded as a bivariate (0,1) variable. As a reference, of the 16 government laboratory technologies in the sample, 6 CRADAs were formed and 3 licenses were signed. For the ten university technologies in the sample, one advanced to the next level of funding. Of the 43 private firm technologies in the sample, 13 advanced to the next level of funding, 6 CRADAs were formed, 2 licenses were signed, and 5 commercial launches were undertaken.

2.2. Data analysis

2.2.1. Artificial neural network model

During the past decade there has been a virtual explosion in research that addresses pattern recognition problems. Particular interest has been paid to artificial neural network (ANN) paradigms. The versatility of neural networks to provide powerful non-linear modeling capabilities without many of the *a priori* assumptions of traditional statistical classifications methods is well-known. For example, traditional modeling methods are based on well-known and convenient *a priori* model forms and related assumptions, such as linearity, normality of error terms, and independence of variables, and can thus suffer from poorly chosen model forms when modeling real data, particularly if these *a priori* assumptions are not characteristic of the underlying process. At best, if model-builders have expertise and applied knowledge of the problem at hand, they may be able to select a model form that reasonably approximates the underlying

generating process. However, poorly chosen model forms can confound the model building process, suffer from inaccurate parameter estimates, produce poor generalization, and thus result in inaccurate interpretations of the underlying process.

Neural networks are, given enough representative data, universal function approximators, and are said to be "data driven". This obviates the need for the initial choice of model form. In a neural network, the model form is approximated concurrently with the parameters through the training process, and, in fact, if the underlying process is indeed linear, neural networks estimation will result in a linear type model. The universal approximation capability of the neural network modeling process is highly adaptive and attractive to many real life problems where the underlying relationships are only vaguely understood, but sample data is abundant. Although ANNs have often been criticized as "black box" treatments, many of the mathematical properties of ANNs have been well established (e.g., Belli, Conti, Crippa, & Turchetti, 1999; Castro, Requena, & Benitez, 2002; Zhang, 2000).

In spite of the potential power of ANN modeling, a number of limitations are also evident. Issues such as overfitting, underfitting and bias are well documented and typical of many non-linear statistical techniques. In addition, while the general functional forms of ANNs are well understood, the nature of neural network estimation, determining initial topologies, developing appropriate training and sampling techniques, and accurate model development and feature selection are not trivial exercises.

Due to these potential limitations with ANN, and to fully investigate the ability of experts to predict technology success, we compare the results of our ANN model with another highly non-restrictive classification model, the Bayesian Data Reduction Algorithm (BDRA) described below.

Since there is significant variance in neural network model building methodology we employed an automated "best practices" model building approach as described by Zhang (2000), similar to that implemented on the Wisconsin Breast Cancer Data database (Hung, Shanker, & Hu, 2002).

We used Netlab (Nabney & Bishop, 2001) to train and evaluate the neural networks for this study. Netlab is a neural network toolbox implemented in Matlab that closely follows the seminal neural network work by Bishop (1995), and doesn't appear to suffer from conversion and other problems reported in other Matlab based neural network algorithms (e.g., Gencay & Selcuk, 2001). Unlike the BDRA, the basic neural network software does not typically aid in model building activities such as feature selection, architecture selection, and global optimization. To address these issues, we developed an automated method paralleling Zhang (2000) and Hung et al. (2002) to perform general model building activities.

In particular, we used a feed-forward, back propagation neural network with logistic activation functions and node biases at each node as implemented by the Netlab mlp function. The error function minimized was SSE using the scaled conjugate gradient algorithm of the netopt function of Netlab. Twenty random starts were used on each architecture trained. Network architectures had exclusively a single hidden layer with between 0 and 3 hidden nodes. The architecture which yielded the best training SSE was retained.

Our feature selection heuristic is a backward elimination method closely following the methodology of Hung et al. (2002). We begin the process with all independent variables in the model. Each variable is then dropped singly, and the best reduced model is chosen to carry forward. Thus, we generate a sequence of models with successively fewer variables, one less variable each time.

Assuming at least one, but not all variables are correlated somehow with the dependent variable, the error along the sequence of successively smaller models follows a common pattern. Early in the sequence the total error of validation and training data begins falling as irrelevant features are removed. Then the total error across successive models becomes fairly constant as the network is able to deal with a moderate level of noise from irrelevant or unnecessary features. Then as truly useful features are eliminated, the total error begins to rise. This pattern is common and recognizable, but unfortunately, is not strictly followed since variability is introduced by local minima in the error function, sample bias, and other factors.

To automate the final selection process we successively reduce the features in the model, and examine the error in the last four models. Four is a subjective number we chose that prevents early selection of a model based on variability in the errors. When the error rises in the last four consecutively smaller models, we end the process, and choose the model with the lowest error in the last four models, i.e., the first of the four. In situations where there are four or fewer independent variables to begin with, we successively reduce the model as described above, and choose the model with the best total error. This follows the spirit of the human expert supervised model selection, in that it chooses a

parsimonious model that performs well on both training and validation data. The pseudo-codes for our modeling technique are available from the authors.

2.2.2. Bayesian Data Reduction Algorithm

The Bayesian Data Reduction Algorithm (Lynch & Willett, 1999, 2003) is a classification method based upon the assumption that the discrete symbol probabilities of each class are *ex-ante* uniformly Dirichlet distributed. It performs feature selection and model building in a highly automated manner, with very few decisions to be made in model building.

In the BDRA, all data are considered to be nominal and symbolically represented as opposed to continuous real values. Furthermore, each symbol is assumed to be initially equally probable. Because not all data are initially represented this way, the BDRA pre-processes the data by discretizing the data into quantiles of a pre-selected size. For instance, if quartiles are selected as the initial level, then the BDRA transforms the real-valued variables within the data into a four level variable (0,1,2,3). The motivation for using discrete values is that it allows for placing a non-informative *prior* on the feature vectors where this would be unproductive for continuous variables. Since an equal number of data points occurs in each quantile, the values can also be considered to be uniformly distributed, satisfying the assumption that they are Dirichlet distributed.

For discrete input data, this process generally has no effect since the initial quantization level can be set equal to the number of discrete categories, but for continuous data the discretizing process may result in some loss of information. However, with the BDRA the quantization level can also be set very high (such as 100 levels) such that it starts to closely approximate continuous data — with the major penalty being computing time. However, previous research also indicates that with the BDRA, there are substantial diminishing returns to higher quantization, and that the incremental increase in the BDRA predictive power is generally highly insignificant beyond a quantization of five or seven, even with continuous data (Kline & Galbraith, 2004; Lynch & Willett, 2003).

Once all variables are represented as nominal values, the algorithm employs a modified backward sequential feature search process for reducing the irrelevant features from the training data of each class. In essence, the algorithm selects those features that provide the best classification performance.

Like ANN classifiers, the BDRA is also data-driven, in that the variables and thresholds on variables are not pre-specified, but determined through the training process. Furthermore, since the BDRA can quantize the data at an arbitrary level of granularity, and any combination of thresholds can be chosen, the solution space can be made up of an arbitrary set of disjoint spaces. This can be viewed as a discrete form of a "universal function approximator".

However, because of its emphasis on feature reduction and quantization, the BDRA differs significantly from ANN classifiers by adjusting the data structure rather than focusing on obtaining the "best fit" model to the existing data. Furthermore, how the feature space is broken into groups is significantly different than linear models and neural network models. Linear models use cutting planes and are centroid-oriented, being based on multivariate normal error assumptions. Neural network models are edge-based, in that separating curves are generated to maximize classification accuracy. The BDRA essentially uses cutting planes perpendicular to dimension axes to separate classes, and have a Bayesian foundation with assumed nominal dimensions.

The BDRA therefore assumes all training and test data as discrete, with the overall cardinality as a Cartesian product of these finite values (see Appendix A). The BDRA then changes the overall cardinality by removal of the features' various thresholds; thus a tertiary feature becomes a binary feature, etc. By utilizing a Combined Bayes Test as the decision function, the BDRA classifies discrete observations given an assumed uniform Dirichlet prior for the symbol probabilities of each class in the training data. Because the Dirichlet prior allows an analytic expression for the probability of error, and has an implicit "penalty term" for model complexity, the feature reduction can be done in a rational and non-heuristic manner (see Lynch & Willett, 1999, 2003).

Essentially the algorithm uses a "greedy" approach to coarsen discrete-valued features, that is the BDRA sequentially iterates through the dimensionality reduction process as opposed to the optimal approach (in the optimal case, all possible ways in which the features can be reduced are determined first, then the probability of error is computed for each possible outcome). The advantage of the BDRA is that it can reduce the quantization complexity of the feature vectors to improve classification performance while simultaneously selecting the relevant features.

Because of its theoretical basis, the BDRA is naturally adept at certain problems such as when the independent variables are almost exclusively nominal. This makes it especially problematic for linear or neural network models which must represent the nominal values as dummy binary variables, thus increasing the dimensionality and

Table 1
Correlation matrix

Variables	EA1	EA2	EA3	EA4	EA5	AGE	EMP	DIVER	EDU	STAGE	FUND	PART
EA1	1.00											
EA2	0.59**	1.00										
EA3	0.62**	0.61**	1.00									
EA4	0.48**	0.62**	0.52**	1.00								
EA5	0.61**	0.67**	0.60**	0.53**	1.00							
AGE	−0.12	−0.07	−0.20**	0.01	−0.24**	1.00						
EMP	−0.26**	−0.23**	−0.37**	−0.14	−0.31**	0.51**	1.00					
DIVER	0.09	0.17*	0.22**	0.17*	0.20*	0.21**	−0.18*	1.00				
EDU	−0.04	0.10	−0.07	0.15*	−0.01	0.34**	0.33**	0.14	1.00			
STAGE	0.19*	0.13	0.22*	0.06	0.06	−0.05	−0.14	0.09	−0.53**	1.00		
FUND	0.19*	0.14	0.04	0.07	0.22**	0.13	0.07	0.22**	0.19*	0.10	1.00	
PART	0.07	0.09	0.01	0.12	0.11	0.18*	−0.21*	0.39**	0.11	0.03	0.32**	1.00

*Prob < 0.10, **prob < 0.05, ***prob < 0.01, two-tailed.

parameterization. Many pure nominal problems are effectively out-of-reach of neural network and linear models. High-dimension problems are also appropriate for the BDRA, since the dimension reduction calculations are very simple.

The BDRA is not to be considered a universally best classifier, but rather to provide superior performance when the "curse of dimensionality" is evident. The BDRA does not necessarily remove whole features, but reduces the quantization complexity by removing discrete levels of the features, one at a time. When additional coarsening of a binary feature is optimal, the feature is removed. Because of these issues, the BDRA should have superior feature reduction characteristic and training time requirements over competitive ANN approaches.

2.2.3. Linear models

As a benchmark methodology a step-wise regression analysis was also performed on the data. While many different linear classifiers are used as benchmarks in neural network comparison studies we felt that the step-wise regression more accurately reflected the balance of model fitting and feature reduction tested in this analysis, particularly since our ANN and BDRA estimations were specifically designed with automated feature reduction characteristics. The threshold probabilities for both variable entry and removals are 0.05. The step-wise linear model is then compared with the ANN and BDRA.

In addition, because of the binary nature of our dependent variable, a final model was estimated using logistic regression. Logistic regression relaxes many of the more restrictive assumptions of OLS, such as normally distributed error terms and strict linearity between the dependent and independent variable. However, when compared to the ANN and BDRA techniques, logistic regression does retain some relatively restrictive assumptions such as independent error terms, low multicollinearlity, and linearity between the independent variables and the logit of the dependent variable.

And while there are some step-wise type logistic techniques that focus on step chi-square outputs, due to the difficulty of measuring model variability with pseudo-R^2s and the problems of developing an automated feature reduction process with logistic regression parallel to the ANN and BDRA, we use the logistic regression more as a confirmatory comparison to our final ANN, BDRA, and step-wise regression models in this study.

3. Results

Table 1 shows the bivariate correlations between the two classes of predictor variables. In general, there seemed to be a relatively high positive internal correlation between the expert assessment variables[4]. However, there appeared to

[4] We also performed a factor analysis on the expert assessment variables. Using a principal component analysis, one factor was extracted (eigenvalue > 1.0) explaining 67.01% of the variance. This could suggest several things. First, that EA1 to EA5 may actually represent items on a single scale; second, given that the expert assessment questions covered very different items, that experts cannot really discriminate in practice between the different dimensions suggested in the technology assessment literature; or third, that the approximately 15 page proposals did not provide sufficient information for the experts to discriminate along the five dimensions.

Table 2
Classification performance: holdout sample comparison

	BDRA mean holdout sample classification rate	ANN mean holdout sample classification rate	LR mean holdout sample classification rate	*T*-test between means, a) BDRA v. ANN b) BDRA v. LR, c) ANN v. LR
Mean correct classification of model using holdout sample, 10 random runs	88.6%	90.2%	77.5%	a) ns b) $p<.05$ c) $p<.05$
Average number of variables in model (both control and expert assessment)	2.0	6.7	6.0	a) $p<.05$ b) $p<.05$ c) ns
Number of "expert assessment" variables	None	None	None	

be significantly lower internal correlation within the class of structural control variables; the exceptions were a 0.53 correlation between EDU and STAGE and a 0.51 correlation between AGE and EMP. In addition, the correlations between the expert assessment variables and the structural control variables appeared relatively low with the exception of the AGE variable, which was significantly and negatively correlated with all of the expert assessment variables (fewer employees viewed more positively by the experts). In sum, the correlation matrix suggests that with the exception of perhaps the employee number variable, the expert assessments appear to be capturing other aspects of the proposal not found in our structural control measurements.

The data was then randomly separated into in-samples (training and validation) and out-of-sample (test or holdout). The test/holdout sample constituted 20%, with the remainder allocated to in-sample. For the BDRA and step-wise regression, the entire in-sample set was used to develop the model. For the ANN, the training set was split into training and validation. Twenty percent of the overall data was used for validation, while 60% of the overall data was used for training. Training and model selection for each method was done completely on the in-sample of the training and validation data sets. The randomized test or "out-of-sample" sample on each multi-sample run was used exclusively for classification rate comparisons, and not for any additional model tuning (Adya & Collopy, 1998; Flexer, 1996; Maier & Dandy, 2000; Prechelt, 1996; Vellido, Lisboa, & Vaughan, 1999).

Comparison studies based upon single sampling are common in the literature, but undesirable due to random influences (Flexer, 1996). The advantages of multi-sampling have long been recognized (Chernick, Murthy, & Nealy, 1985; Jain, Duin, & Mao, 2000). Therefore we use a ten sample randomized multi-sampling process in this study. In addition, there have been serious questions raised about many of the published applications of estimated neural network models. Prechelt (1996), for example, found only 22% of published studies met his minimum criteria for modeling, while Flexer (1996) found less than 10% of published papers use a completely independent test set. Adya and Collopy (1998), Vellido et al. (1999), and Maier and Dandy (2000) have all expressed similar concerns about the quality of modern ANN modeling practice. In our analysis and results reporting we attempt to follow best practices as described by the above authors.

3.1. Classification rate

Mean and standard deviation based upon the ten randomized samples were computed. Mean differences between the BDRA and ANN methods for the test-sample were examined with the *t*-statistic based upon a paired sample analysis (see Looney, 1988). Table 2 presents the test/holdout sample classification rate performance for the BDRA, ANN, and step-wise linear regression.

In developing a predictive model of technology success, both the neural network and the Bayesian Data Reduction Algorithm significantly outperformed the linear model. The ANN models resulted in an average 90.2% correct classification in the holdout sample. However, it should be noted that the in-sample correct classification was 97.6%, which suggests a possible overfitting problem for the ANN model. The BDRA results in a similar 88.6% correct classification, however no overfitting was evident with the BDRA. This compares to a 77.5% correct classification in the holdout sample for the step-wise linear regression.

Table 3
Logistic regression model

Variable	Structural coefficient	Full specified coefficient	Expert scale coefficient
Constant	−4.750	−3.574	−3.741
AGE	2.122***	2.177***	2.181***
EMP	−1.674***	−1.802***	−1.749***
DIVER	−1.797**	−1.693**	−1.761**
EDU	0.476	0.520	0.517
STAGE	1.167***	1.216***	1.191***
FUND	−0.792**	−0.901**	−0.812**
PART	0.988*	0.898	0.902
EA1	–	0.151	–
EA2	–	0.046	–
EA3	–	−0.116	–
EA4	–	−0.082	–
EA5	–	−0.189	–
EXPERT SCALE	–	–	−0.031
Nagelkerke R^2	0.587	0.601	0.591
N	180	180	180

*Prob < 0.10, **prob < 0.05, ***prob < 0.01.

When examining the contribution of expert assessments to the predictive model, in all three cases, none of the five expert assessment variables entered into the model when the control variables where also included. In other words, the model estimation and correct classification rates were achieved purely by using the seven control variables of age of firm, number of employees, education level of technology team, diversification, total previous funding for technology, partnerships, and stage of technology development.

3.2. Feature reduction

The BDRA is not specifically designed to optimize classification only, but rather to achieve a balance between reducing feature complexity and classification performance. In essence the BDRA attempts to achieve a significant reduction in both the number of dimensions and their quantization complexity while still achieving acceptable classification rates. The number of features (variables) in the final model is also presented in Table 1.

As expected the BDRA significantly reduces the predictive model of technology success to a lower level of dimensionality than both the ANN and LR. In all 10 sample runs, the final BDRA model used only two control variables, stage of technology development (later stages resulted in greater success) and size of firm (smaller firms resulted in greater success) to achieve its correct classification of 88.6%. Depending on the sample run, the ANN used between 6 or 7 control variables in the final model. The step-wise linear regression consistently resulted in six control variables as a final model; only the "partnership" variable was excluded in the LR model.

3.3. Logistic regression

As a final analysis, we were interested in further confirming if the expert reviewers added any additional explanatory value above that provided by the organizational control model. Table 3 presents this analysis. First, a fully specified model, including both control and expert reviewer assessment variables was analyzed. In the fully specified model none of the expert reviewer variables were statistically significant in the model, and were excluded from the final model estimation. Table 3 also presents a final, parsimonious logistic regression model (excluding the non-significant expert assessment variables). Similar to the step-wise regression model which used six structural control variables, the final logistic regression resulted in a predictive model utilizing six of the structural control variables.

In addition, since the five expert assessment variables were highly correlated (Table 1) we also performed a factor analysis on the expert assessment variables. Using a principal component analysis, one factor was extracted (eigenvalue > 1.0) explaining 67.01% of the variance. We then created a simple additive scale (EXPERT SCALE) of the five expert assessment items, similar to what is typically done in early stage technology evaluations. A logistic

regression was then performed using the seven structural control variables and the aggregate expert assessment scale (Table 3). Again, the expert assessment scale was not significant, while five of the structural control variables were statistically significant. Overall, the three logistic regression models appeared highly stable, and the addition of the expert assessment variables increased the Nagelkerke R^2 by only 0.014 (including all five expert assessment variables) and 0.004 (including the expert scale variable). This again appears to confirm the results from our less restrictive BDRA and ANN models.

Overall, the results of the neural network analysis, the Bayesian Data Reduction Algorithm, the step-wise regression, and the logistic regression all suggest that the incremental value of expert reviewers to predict future technology success beyond a basic control or structural model for early to mid-stage technologies may be relatively small.

4. Conclusions

Predicting the future successes of early to mid-stage technologies is one of the most fundamental of all technology management functions. Grant funding agencies, university technology transfer offices, corporate R&D departments, nascent entrepreneurs, and venture capitalists all partake in this exercise. But, while we know, at least in hindsight, some of the factors that appear to be related to successful technological commercialization (e.g., a superior product), we know far less about the precursors to technology success and whether or not these factors can actually be predicted.

This study attempts to examine some of these issues. Using a sample of sixty-nine early to mid-stage homeland defense technologies, we develop an *ex-ante* predictive model of future technology success.

Several conclusions are suggested by the analysis. First, a reasonably good predictive model can be constructed from the control, or organizational factors; that is, the characteristics of the firm, the business model, the technology, and the research team. This finding somewhat parallels Roure and Keeley's (1990) argument that there is usually sufficient information in a written business plan to predict the success of a technology-based venture as well as more recent research that suggest that reasonable predictive models can be constructed using various organizational, structural, and human capital variables (e.g., Baum & Silverman, 2004). Along that line, in our study, we found that sufficient information can be contained in a short "business plan-like" technology application to develop a statistically significant model of technology success.

From this analysis it appears that focused, small and more established firms/labs with mid-stage technologies (lab prototype or later) are associated with future technology successes. There is additional support that higher education levels of the research team and formal strategic alliances also contribute to future success. Our analysis also indicates that financially leaner might be better, with more fully funded technologies actually underperforming those with less early stage external funding. This apparently successful combination of small, focused and financially lean operations may create an overall "lean enterprise" where non-value added elements are continuously scrutinized and eliminated (e.g., Panchak, 2003; Rossetto & Franceschini, 1995).

Second, there is high consistency between our linear models (logistic and step-wise) and our model developed using a neural network approach. This consistency provides a degree of confirmatory confidence to our analysis. In addition, the BDRA provided outstanding correct classification of technology success using only two control variables, whereas the final estimated models using neural networks, step-wise regression, and logistic regression used between 6 and 7 control variables.

Third, on the average, expert reviewers provide little or no predictive value given the data in our study. As a group, the expertise of venture capitalists, scientists, engineers, technology transfer officers, corporate and program R&D managers, and technology commercialization consultants added no statistically significant value to our predictive models. This conclusion is consistent across all four techniques used in the study. Given that most grant proposals, business plans, and internal R&D requests at large corporations are initially screened by a panel of "experts" based upon a formal review of a written proposal or document, our finding appears to challenge the overall validity of this process. Perhaps so-called "experts", in general, are unable to really assess the overall commercial potential of early stage technologies, or perhaps there are identifiable characteristics of certain people that provide greater expertise in predicting future success than others. As a group, however, our initial analysis of an actual panel of experts reviewing actual proposals appeared to indicate little or no predictive power. Certainly, the overall validity of early stage screening by experts needs further investigation.

While the results of our study are intriguing, there are also several limitations. First, because we studied only CCAT funded technologies, there may be a selection bias in the sample. Over 500 technology proposals were not funded by CCAT, and these were not included in the present sample. However, we believe that there is sufficient variation in both the reviewer evaluations and the organizational control variables to reasonably examine the appropriate factors, and to move toward developing a predictive model of technology success. For example, each of the five reviewer assessment items had scores ranging between a low of 0 and a high of 10, with variances ranging between a low of 3.71 and a high of 6.20; the organizational variables also ranged between the high and low points on each scale.

Second, although our analysis indicated, at best, only a marginal predictive ability by the expert reviewers, the reviewer scores in this study were based solely upon their reading of a lengthy proposal and summary of the firm, team members, target market, and technology. While in some situations this approaches reality (such as initial government grant reviews, large scale technology transfer office decisions, and preliminary equity investor business plan screenings), usually interested investors, consultants, and managers would ultimately get richer information through personal interviews and due diligence prior to a final assessment of a technology's chances. During a due diligence process, a broader set of questions are answered than are usually found in initial proposals or business plans. And during a personal presentation of the technology proposal, an expert panel can investigate in far more detail questions they might have about the technology, team, or market potential. Thus, while the findings of this study perhaps challenge the validity of expert panel assessments as an initial screen for written proposals, it might be that the same experts, because of their very expertise, can extract far more relevant information during a due diligence investigation or by asking questions during personal interviews or presentations.

In spite of these recognized limitations, we believe that our results raise important issues regarding the capability of developing predictive models, and the interplay between organizational factors and expert knowledge. In this spirit, we call for more aggressive empirical testing of the predictive capability of expert reviewers, technology "readiness" scoring models, and stage-gate type technology development programs.

Appendix A. Technical discussion of Bayesian data reduction technique

Unlike most ANN approaches which tune the model to fit unquantized feature data, the BDRA searches for the optimal quantization complexity and feature space that minimizes the probability of error based on the following formula

$$P(e|X) = \sum_{k=1}^{C} \sum_{y} \sum_{X} P(H_k) I_{\{f_k \le f_l,\ \text{for all } k \ne l\}} f_k \tag{1}$$

where, $f_k = f(y|x, H_k) = \frac{N_y!(N+M-1)!}{(N+N_y+M-1)!} \prod_{j \in H_k} \frac{(x_j+y_j)!}{x_j! y_j!}$; M equal to total discrete variables and C is the total number of classes with $k \in \{1, \ldots C\}$; and $j \in H_k$ is defined as all discrete symbols, j, associated with class k, and with the class-labeling feature is equal to k; H_k is the hypothesis defined as $p_y = p_{\text{all } j \in H_k}$, and $\left\{\sum_{k=1}^{C} \sum_{j \in H_k}^{M} p_j = 1\right\}, X$ is the entire collection of training data from all C classes; $x_{j \in H_k}$ is the number of occurrences of the jth symbol in the training data defined for all $j \in H_k$; $N\{N = \sum_{j=1}^{M} x_j\}$ is the total number of training data, where the fraction belonging to the kth class is given by, $\{N_k = \sum_{j \in H_k} x_j\}$ y_j is the number of occurrences of the jth symbol in the test data; $N_y\{N_y = \sum_{j=1}^{M} y_j\}$ is the total number of test data; $I_{\{x\}}$ is the indicator function. Note that a typical situation in practice (versus a multiple-sample research study) usually involves only one observation of test data (i.e., $N_y = 1$) In this case, $f(y|x,H_k)$ of Eq. (1) becomes, $f(y_i = 1|x, H_k) = \frac{x_{j \in H_k}+1}{N+M}$.

References

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, *17*, 481−495.

Ajamian, G., & Koen, P. A. (2002). Technology stage gate: A structured process for managing high risk, new technology projects. In P. Belliveau, A. Griffin, & S. Sorermeyer (Eds.), *PDMA toolbook for new product development* (pp. 267−295). New York: John Wiley and Sons.

Astebro, T. (2004). Key success factors for technological entrepreneurs' R&D projects. *IEEE Transactions on Engineering Management*, *51*(3), 314−321.

Balachandra, R., & Friar, J. (1997). Factors for success in R&D projects and new product innovation: A contextual framework. *IEEE Transactions on Engineering Management*, *44*, 276−287.

Baum, J., & Silverman, B. (2004). Picking winners or making them? Alliances, patents and human capital as selection criteria in venture financing and performance of biotechnology start-ups. *Journal of Business Venturing*, *19*, 411−436.

Belli, M., Conti, M., Crippa, P., & Turchetti, C. (1999). Artificial neural networks as approximators of stochastic processes. *Neural Networks*, *12*, 647−658.

Belliveau, P., Griffin, A., & Somermeyer, S. (Eds.). (2002). *The PDMA toolbook for new product development* New York: John Wiley and Sons.

Bishop, C. (1995). *Neural networks for pattern recognition.* Oxford: Clarendon Press.

Carland, J., Hoy, F., Boulton, W., & Carland, J. (1984). Differentiating entrepreneurs from small business owners: A conceptualization. *Academy of Management Review*, *9*, 354−359.

Castro, J., Requena, I., & Benitez, J. (2002). Interpretation of artificial neural networks by means of fuzzy rules. *IEEE Transactions on Neural Networks*, *13*, 101−116.

Chernick, M., Murthy, V., & Nealy, C. (1985). Application of bootstrap and other resampling techniques: Evaluation of classifier performance. *Pattern Recognition Letters*, *3*, 167−178.

Christensen-Szalanski, J., & Fobian, C. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*, 147−168.

Cooper, R. (1979). The dimensions of industrial new product success and failure. *Journal of Marketing*, *43*, 93−103.

Cooper, R. (1993). *Winning at new products* (2nd ed.) Redding, MA: Addison–Wesley.

Cooper, R. (1998). *Product leadership: Creating and launching superior new products.* Reading, MA: Perseus Publishing.

Cooper, R. (2001). *Winning at new products.* Reading, MA: Perseus Publishing.

Cooper, R., Edgett, S., & Kleinschmidt, E. (2002). Optimizing the stage-gate process: What best practice companies are doing- Part I. *Research Technology Management*, *45*(5), 21−27.

Davidsson, P., & Honig, B. (2003). The role of social and human capital among nascent entrepreneurs. *Journal of Business Venturing*, *18*, 301−331.

Eldred, E., & McGrath, M. (1997). Commercializing new technology-II. *Research Technology Management*, *40*(2), 29−33.

Eldred, E., & Shapiro, A. (1996). Technology management. In M. E. McGrath (Ed.), *Setting the PACE in product development* Boston, MA: Butterworth and Heinemann.

Elton, E., Martin, J., & Blake, C. (1996). Survivorship bias and mutual fund performance. *Review of Financial Studies*, *9*, 1097−1120.

Flexer, A. (1996). Statistical evaluation of neural network experiments: Minimum requirements and current practice. In R. Trappl (Ed.), *Proceedings of the Thirteenth European Meeting on Cybernetics and Systems Research*, *Vol. 13*. (pp. 1005−1008).

Gencay, R., & Selcuk, F. (2001). Software reviews: Neural network toolbox 3.0 for use with MATLAB. *International Journal of Forecasting*, *17*, 305−322.

Griffin, A. (1997). Modeling and measuring product development cycle time across industries. *Journal of Engineering and Technology Management*, *14*, 1−24.

Heslop, L., McGregor, E., & Griffith, M. (2001). Development of a technology readiness assessment measure: The cloverleaf model of technology transfer. *Journal of Technology Transfer*, *26*, 369−384.

Hung, M., Shanker, M., & Hu, M. (2002). Estimating breast cancer risks using neural networks. *Journal of the Operational Research Society*, *53*, 222−231.

Jain, A., Duin, R., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4−37.

Kline, D., & Galbraith, C. (2004). Performance analysis of the Bayesian Data Reduction Algorithm (BDRA) in two group dependent variable data. *Technical Report 04-01Center for Commercialization of Advanced Technology, San Diego, CA*.

Koen, P., Ajamian, G., Boyce, S., Clamen, A., Fisher, E., Fountoulakis, S., et al. (2002). Fuzzy-front end: Effective methods, tools and techniques. In P. Belliveau, A. Griffin, & S. Sorermeyer (Eds.), *PDMA Toolbook for new product development* (pp. 2−35). New York: John Wiley and Sons.

Linton, J., Walsh, S., & Morabito, J. (2002). Analysis, ranking and selection of R&D projects in a portfolio. *R&D Management*, *32*(32), 139−148.

Looney, S. (1988). A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, *8*, 5−9.

Lynch, R., & Willett, P. (1999). Performance considerations for a combined information classification test using Dirichlet priors. *IEEE Transactions on Signal Processing*, *47*, 1711−1715.

Lynch, R., & Willett, P. (2003). Bayesian classification and feature reduction using uniform Dirichlet priors. *IEEE Transactions on Systems, Man, and Cybernetics*, *33*(3), 448−464.

Maidique, M., & Zirger, B. (1984). A study of the success and failure in product innovation: The case of the U.S. electronics industry. *IEEE Transactions on Engineering Management*, *31*, 192−203.

Maier, H., & Dandy, G. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modeling issues and applications. *Environmental Modelling and Software*, *15*, 101−124.

McCollough, R. (1998). The ProGrid system: A customized expert system for evaluation of intellectual capital. *Proceedings of the workshop on evaluating technology for commercial exploitation* Ottawa: Federal Partners for Technology Transfer.

McGrath, M. (1996). *Setting the PACE in product development.* Boston, MA: Butterworth-Heineman.

Mock, J., Kenkeremath, D., & Janis, F. (1993). *Moving R&D to the marketplace: A guidebook for technology transfer managers.* Falls Church, VA: Technology Prospects, Inc.

Montoya-Weiss, M., & Calantone, R. (1994). Determinants of new product performance: A review and meta-analysis. *Journal of Product Innovation Management*, *11*, 397−417.

Nabney, I., & Bishop, C. (2001). *Netlab neural network software.* http://www.ncrg.aston.ac.uk/netlab/

Ozer, M. (1999). A survey of new product evaluation models. *Journal of Product Innovation Management*, *16*, 77−94.

Panchak, P. (2003). Pella drives lean throughout the enterprise. *Industry Week*, *252*(6), 74−75.
Prechelt, L. (1996). A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. *Neural Networks*, *9*, 457−462.
Rossetto, S., & Franceschini, F. (1995). Quality and innovation: A conceptual model of their interaction. *Total Quality Management*, *6*(3), 221−229.
Roure, J., & Keeley, R. (1990). Predictors of success in new technology-based ventures. *Journal of Business Venturing*, *5*, 201−220.
Schoenborn, T., & Fredette, J. (1997). *So you want to set up a technology evaluation program?* Cherry Hill, NJ: Federal Laboratory Consortium.
Shepard, D., & Zacharakis, A. (2002). VCs decision processes: Evidence suggesting more experience may not always be better. *Journal of Business Venturing*, *18*, 381−401.
Shane, S., & Stuart, T. (2002). Organizational endowments and the performance of university start-ups. *Management Science*, *48*(1), 154−170.
Vellido, A., Lisboa, P., & Vaughan, J. (1999). Neural networks in business: A survey of applications. *Expert Systems with Applications*, *17*, 51−70.
Zacharakis, A., & Meyers, G. (2000). The Potential for actuarial decision models: Can they improve the venture capital investment decision? *Journal of Business Venturing*, *15*, 323−346.
Zacharakis, A., & Shepard, D. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, *16*, 311−332.
Zhang, G. (2000). Neural networks for classification: A survey. IEEE Transactions on Systems. *Man, and Cybernetics - Part C*, *30*, 451−462.