Economics Bulletin

Volume 0, Issue 0

Skills for the future – forecasting firm competitiveness using machine learning methods and employer–employee register data

Pål Børing
NIFU Nordic Institute for Studies innovation, research and education

Arne Martin Fevolden
NIFU Nordic Institute for Studies innovation, research and education

André Lynum
Tidal Music AS

Abstract

This article investigates whether skills data can be used to forecast firm competitiveness. It makes use of an employer–employee register dataset consisting of detailed information about the educational background of all employees in the manufacturing sector in Norway and uses this data to predict the manufacturing firms' revenues five years into the future. The predictions are carried out by employing three machine learning models – lasso regression, random forest and gradient boosting. The results show that machine learning models using skills data can provide reasonably good forecasts of firm competitiveness. However, the results also show that these models become less reliable at the "extreme ends" and that they predicted extreme increases or decreases in revenues poorly.

Citation: Pål Børing and Arne Martin Fevolden and André Lynum, (2021) "Skills for the future – forecasting firm competitiveness using machine learning methods and employer–employee register data", *Economics Bulletin*, Vol. 0 No. 0 p.A60.
Contact: Pål Børing - paal.boring@nifu.no, Arne Martin Fevolden - arne.fevolden@nifu.no, André Lynum - andre.lynum@tidal.com.
Submitted: December 07, 2020. **Published:** April 09, 2021.

1. Introduction

Skills serve as the foundation for both the commercial success of firms and the prosperity of countries. At the macroeconomic level, skills shortages can act as barriers for economic growth in developing countries, and skills mismatches can be a source of productivity loss in developed countries (Brunello and Wruuck 2019). At the microeconomic level, employee skills are regarded by many as a firm's main sources of competitive advantage, and movement of skilled workers between related industries are associated with rapid rates of technological development and innovation (Boschma et al. 2008). Given that skills function as an important catalyst for such a diverse set of economic activities, it is not surprising that skills features in the explanation for a wide range of economic phenomena. However, it is somewhat surprising that skills have yet to feature strongly in an equally important analytical task – namely, prediction. To the extent of our knowledge, there are few if any studies that have used skills in forecasting, and the aim of this article is to help remedy this situation by investigating whether data on employee skills can be used to forecast the competitiveness of firms in the Norwegian manufacturing sector.

Economic forecasting has been a regular task for economic professionals since (at least) the early 1900s. Economic professionals have been tasked with predicting everything from economic growth of large and powerful countries to the to the return on investments from minority positions in small, startup firms. Nevertheless, this relatively old task has recently undergone something of a transformation, as a new class of methods have been adopted from computer science and computational statistics – called machine learning. These machine learning methods can – according to their proponents – produce more accurate predictions than traditional econometric models (especially in the extreme ends of the distribution) and can enable researchers to make use of unconventional datasets that traditional econometric methods would struggle to make sense of. However – as their opponents point out – these machine learning models pay a price for these abilities: they are not open to interpretation in the same way as econometric models are and cannot make use of the same type of statistical tests to quantify the uncertainty of their prediction (Mullainathan et al. 2017).

In this article, we will make use of machine learning methods partly because we want to make use of an unconventional dataset and partly because we want to make more accurate predictions. We want to use of a dataset which is both sparse and high dimensional – consisting of 280 different educational variables of which most of the firms only need a few to classify their employees. We also want to make predictions that are not only good on average, but that are also good at the extreme ends – that is, we want to be able to make predictions that are accurate also for firms that experience a radical change in fortune and that transition from a low to a high level of competitiveness, or vice versa.

But what is it with skills that should enable us to forecast competitiveness? There are two different views in the business and economics literature on how employee skills contribute to firm competitiveness. The first view – which we can refer to as the *stock* perspective – maintains that competencies (or capabilities) are the foundation for a firm's competitiveness and that these competencies are created when a firm successfully combines employee skills with organizational structures and routines (Prahalad and Hamel 1990; Teece 1997). The stock perspective is static in the sense that its focus is mainly on the firm's current inventory of skills and competences. The second view – which we can refer to as the *flow* perspective – claims that firms can improve their competitiveness by hiring employees from different firms or sectors that bring with them new skills and can spur innovation and change. The flow

perspective is dynamic in the sense that its focus is on changes in the firm's skills base (Boschma et al. 2008). In this article, both stock and flow perspectives are relevant when we try to forecast competitiveness. More specifically, our research questions can be summarized as the following:

(i) Can information about the stock and flow of employee skills be used to accurately predict a firm's future competitiveness?

(ii) Are the predictions accurate for firms that experience a dramatic change in competitiveness?

The article will answer these two questions by employing three machine learning methods – lasso regression, random forest and gradient boosting. The article will first provide an answer to whether data on employee skills can be used to accurately predict a firm's future competitiveness. This will be done by estimating a baseline model that predicts future revenues based on past revenues and explore whether the baseline model's prediction can be improved by adding skills data and using machine learning methods. The article will then provide an answer to whether this prediction is accurate for firms that experience a dramatic change in fortune (by transitioning from a low to a high or a high to a low level of competitiveness). This will be done by dividing the dataset into 10 quantiles based on how much the firms' competitiveness decreases/increases, within the chosen time interval, and exploring how accurate the models predict the competitiveness of the firms that fall in the lower and upper quantiles. In addition to the baseline and machine learning models, a multiple regression model is also estimated to illustrate the difficulties using sparse and high dimensional data with traditional econometric models.

2. Forecasting Design and Data

In this article, we want to use information about employee skills to predict competitiveness. However, both skills and competitiveness are concepts that are vague and abstract and that must be operationalized before they can be employed in a quantitative analysis.

Competitiveness is a term that is loosely defined as a firm's ability to succeed in the international markets and can encompass a number of firm characteristics, such as market share, profits, turnover and productivity (Bhawsar and Chattopadhyay 2015). In our analysis, we will only make use of one these characteristics – namely, **turnover**. We will, accordingly, view a high turnover as an indicator of high competitiveness and an increase in turnover as an indicator of improved competitiveness. It is the firms that increase their turnover that we label "winners".

Skills is a term that has been widely used in the social sciences and has been defined in many different ways. In our analysis, we will define skills simply as the **employee's educational background**. Accordingly, a firm's stock of skills is the number of employees broken down according to the employees' educational background and the flow of skills is the change in a firm's stock of skills from one time period to another.

We will employ a **forecasting design** where we in year **t** predict a firm's turnover five years into the future, in year **t+5**. This prediction is based on **(1)** the firm's turnover in year **t**, **(2)** the firm's employee skills in year **t**, which represents the present stock of skills, and **(3)** the change in employee's skills from year **t-1** to **t**, which represent the flow of skills (see Table I). Selecting five years as the forecasting target is somewhat arbitrary. Three or ten years could have been equally good choices. In a more comprehensive analysis, it would be natural to employ a series of targets and compare them. Nevertheless, the brevity of the Economics Bulletin format calls for a less extensive analysis, and five years were chosen as a good, "middle-of-the-road" forecasting target.

2.1 The Data

The empirical analysis is based on matched employer–employee register data of Norwegian manufacturing firms, which comprises annual administrative files for the years 2010, 2011 and 2016 from Statistics Norway. The data contains yearly information on highest attained education among all employees, and on turnover among all plants, in the manufacturing sector in Norway for these years. In the analysis, firms are defined at the plant level.
The sample of firms consists of all firms that are registered with non-missing turnover in the manufacturing sector both in 2011 and 2016. In addition, we only include firms in the sample where there are registered employees in each firm both in 2010 and 2011. There are 7029 unique manufacturing firms in the sample.

Table I: Predictors and Target Variables

	Year	Competitiveness (firm turnover)	Skills (employee education)
Predictors	2010		280 features
	2011	1 feature	280 features
Target	2016	1 feature	

2.2 Target and Predictor Variables

The **target variable** is turnover in 2016. Turnover is the sum of payment of sales to customers, sales of goods for resale, and gross income from other business activity. Turnover includes income from rent and commission income, but not government subsidies or profit from the disposal of fixed assets. Value added tax is not included in the turnover either. We measure turnover in millions NOK, which is about $114,000 US dollars at the time of writing.

We make use of three sets of **predictor variables**. The first set consists of only one variable, turnover in 2011, which like the target variable is measured in millions NOK. The second set consists of 280 educational variables based on employees' educational background. An employee's highest attained education is used as information on his or her educational background. Education is based on the Norwegian Standard Classification of Education

(NUS2000). This is a 6-digit code system that classifies educational activities by level and subject. We use the first three digits of NUS2000 in order to compute the 280 educational variables, which classify the employees by different professional groups. Each of these variables measures the number of employees in a firm that has completed a particular three-digit code of NUS2000 **in 2011**. The third set of predictors consists of an equivalent set of 280 three-digit codes of NUS2000, but these variables measure the change in the number of employees with a particular educational background from **2010** to **2011**. In total, we use 561 predictor variables. When we carried out the predictions using the Lasso model, we standardized the predictor variables (mean zero and standard deviation one). Otherwise, the predictors were used as is.

3. Econometric and Machine Learning Models

We use five different models to forecast firm turnover – one baseline model, one naïve econometric model and three machine learning models. The first two models are primarily of instrumental value. The baseline model is instrumental in the sense that it establishes a performance threshold that machine learning models are measured against. If one of the machine learning models fail to beat the baseline predictions, we might assume that *that* model type performs poorly on the skills data; if all machine learning models fail, however, then we might assume that the skills data might lack predictive power. The naïve econometric model is instrumental in the sense that it is used to test how a traditional econometric model does on high-dimensional, sparce data and justifies the use of machine learning models for forecasting.

Our main interest lies is the machine learning models and their ability to use skills data to forecast turnover. The reason why we use three different machine learning models (and not only one) is that it is almost impossible to tell in advance which machine learning model will perform best on a particular dataset (a fact that is often referred to as an inductive bias) (Kuhn and Johnson 2013: Ch. 4). It is therefore common practice to explore several types of machine learning models and compare them according to appropriate metrics. In addition, such models also differ in terms of interpretability and computational complexity. It is therefore interesting to see how well different types of machine learning models perform on a dataset, since we might later want to switch between models to trade accuracy for simplicity or interpretability.

3.1 Baseline Model

We use a univariate linear regression – with firm turnover in 2016 as the target variable and firm turnover in 2011 as the predictor variable – as our baseline model. This baseline model can be written in the following form:

$$(1) \quad y_i = \alpha_0 + \alpha_1 x_{1i} + \varepsilon_i,$$

where y_i represents the turnover in 2016 for the *i*th firm, α_0 is the intercept, α_1 is the coefficient for the variable 'turnover in 2011', x_{i1} is the *i*th firm's turnover in 2011 and ε_i is an error term. The addition of skills to the baseline model is expected to provide significantly improved predictive performance.

3.2 Naïve Econometric Model – Multiple Linear Regression

We will use multiple linear regression – with firm turnover in 2016 as the target variable and firm turnover for 2011 and all the skills variables for 2010 and 2011 as the predictor variables – as our naïve econometric model. This naïve econometric model can be written in the following form:

$$(2)\quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_2 x_{281i} + \beta_{282} x_{282i} + \cdots + \beta_{561} x_{561i} + u_i,$$

where $\beta_2, \ldots, \beta_{281}$ are the coefficients for the skills variables for 2011, $x_{i2}, \ldots, x_{i281}$ are the number of employees with the particular skillset for the *i*th firm in 2011, $\beta_{282}, \ldots, \beta_{561}$ are the coefficients for the skills variables for 2010, $x_{i282}, \ldots, x_{i561}$ are the number of employees with the particular skillset for the *i*th firm in 2010, and u_i is an error term. We describe this model as a naïve, since no attempt has been made to transform the data or reduce the number of features to avoid complications related to collinearity or sparsity.

3.3 Machine Learning – Lasso

The first machine learning model we will use is Lasso (least absolute shrinkage and selection operator). Lasso is a linear regression method that adds a penalty to non-zero regression coefficients. Like OLS, it estimates the parameters by minimizing the sum of square errors (SSE), but unlike OLS, it adds an additional penalty to the equation, which is equal to the sum of the absolute values of the coefficients (γ_j) multiplied by a constant (λ), which acts as a hyperparameter that must be set appropriately:

$$(3)\quad \mathrm{SSE}_{L1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\gamma_j|.$$

This penalty – often referred to as L1 norm or L1 regularization – can shrink and remove coefficients. Since shrinking and removing coefficients can reduce variance without a substantial increase in bias, L1 regularization tend to improve the model's adaptation to new data. This is especially true when the data is sparse, exhibits widespread multi-collinearity and contains a large number of features compared to observations (Kuhn & Johnson 2013: ch. 6; Hastie et al. 2015: ch. 1 & 2) – which is the case for our dataset. How effective the L1 regularization is at shrinking coefficients and improving model accuracy depends on the tuning of the hyperparameter λ. A high λ will increase the shrinkage factor and may initially improve predictive performance due to regularization. However, as λ is increased further, the model will at some point degenerate into the dataset mean. To achieve the best predictive performance, this hyper parameter must be adapted to the dataset. We fine-tuned this hyper parameter by exploring a range of suitable values using five-fold cross-validation, repeating this process with finer granularity until a near optimal value had been identified (as suggested by Kuhn & Johnson 2013: ch. 4).

3.4 Machine Learning – Random Forest and Gradient Boosting

The two other machine learning models that we use are two regression-tree based models – Random Forest and Gradient Boosting. These two models belong to a group of machine learning methods called ensemble models. Ensemble models combine together the predictions from multiple machine learning models with weak predictive performance. The rational for

these models is that there are some classes of machine learning methods – such as shallow decision tree models – where individual models tend to overfit or otherwise have weak performance. However, learning theory has shown that the predictive performance can become competitive when many such models are combined in an ensemble prediction. What distinguishes Random Forest from Gradient Boosting is the way the model is trained and assembled. Random Forest employs a bagging (bootstrap aggregation) technique where each tree is trained using a bootstrap sample from the training data, and each split in each tree is found by searching random subsets of features. The resulting group of trees gives us the ensembling improvement described above. Gradient Boosting creates a series of decision trees in sequence where each new decision tree is trained on the residual error from the combination of the previously learned decision trees. As with Random Forest, this method produces a group of trees that as an ensemble constitutes a more accurate model (James et al. 2013: ch. 8) These two ensemble models have many advantages due to their decision tree foundation. For example, they are usable with sparse or even missing data and are insensitive to variation in the scale of the features. This combined with their tendency for strong predictive performance has made these models very popular among machine learning practitioners. But like the Lasso method there are hyper parameters that must adapted to the dataset. In our experiments the hyper parameter values are set using the grid search method implemented in Python's SciKit-learn library.

3.5 Model Validation and Evaluation

We have followed the suggestions from Kuhn and Johnson (2013: ch. 4) to validate our models using cross-validation (rather than train/test-split) and use multiple performance metrics to evaluate the models. More specifically, we have chosen to use a fivefold cross validation and employ three performance metrics – Mean square error (MSE), mean absolute error (MAE) and R^2.

4. The Empirical Results

In order to investigate information about how the stock and flow of employee skills will be used to accurately predict a firm's future turnover, we made use of five different models: a baseline model, a naïve econometric model and three machine learning models – Lasso, Random Forest and Gradient Boosting. Table II provides an overview of how well each of the models performed along the three metrics MAE, MSE and R^2.

Table II: Model Accuracy Metrics

	Mean Absolute Error	Mean Square Error	R-squared
Baseline	61.28	56757.47	0.67
Multivariate Linear Regression	3559458.91	89020197503719520	-761661674909.91
Lasso Regression	54.48	42543.38	0.74
Random Forest	37.46	37930.71	0.77
Gradient Boosting	38.99	35551.86	0.78

As we can see from Table II, our baseline model received a reasonably high R^2 score of 0.67. This implies that we can make a fairly good prediction for firm's turnover 5 years into the future by assuming that it is identical to the present turnover, plus or minus an adjustment factor ($\alpha_0 + \alpha_1 x_{1i}$) that can account for inflation and sector growth. Despite the relatively high performance of the baseline model, we can see that all the machine learning models did significantly better. The Lasso model had an R^2 of 0.74 and Random Forest and Gradient Boosting had an R^2 of 0.77 and 0.78 respectively. Similar results can also be seen for MAE and MSE, which are considerably lower for the machine learning models than for the baseline model. These results indicate that we can indeed significantly improve our predictions of a firm's turnover by including data on employee skills. However, the results also indicate how much these skills data contribute to improving our predictions depend on the choice of model.

If we look at the performance of the Naïve econometric model, we see that it fails to produce any reasonable predictions from the skills data. It performs far worse than the baseline model, and its R^2 is high and negative, which tells us that the model's predictions are far less accurate than assuming that all firms' turnover will be equal to the mean turnover in 2016 (target value mean). There are many possible explanations for the model's poor performance, among others, potential collinearity between the skills variables for different years. Nevertheless, the naïve econometric model illustrates that it is far from trivial to make predictions from the skills data.

If we look at the performance of the machine learning models, we can see that the two ensemble models perform best. The Random Forest and the Gradient Boosting have an R^2 of 0.77 and 0.78 respectively. However, the Random Forest model has a higher MSE and a lower MAE than the Gradient Boosting model. This indicate that the Random Forest is slightly more accurate than the Gradient Boosting model for most predictions but has some "big misses" that the MSE accentuates up by squaring the error function. The Lasso model is not far behind the ensemble models in terms of accuracy. Given that Lasso is a simpler model to train and interpret, it might be a good choice in circumstances when computational simplicity and interpretability is important. When accuracy is most important, the ensemble models are the preferred choice.

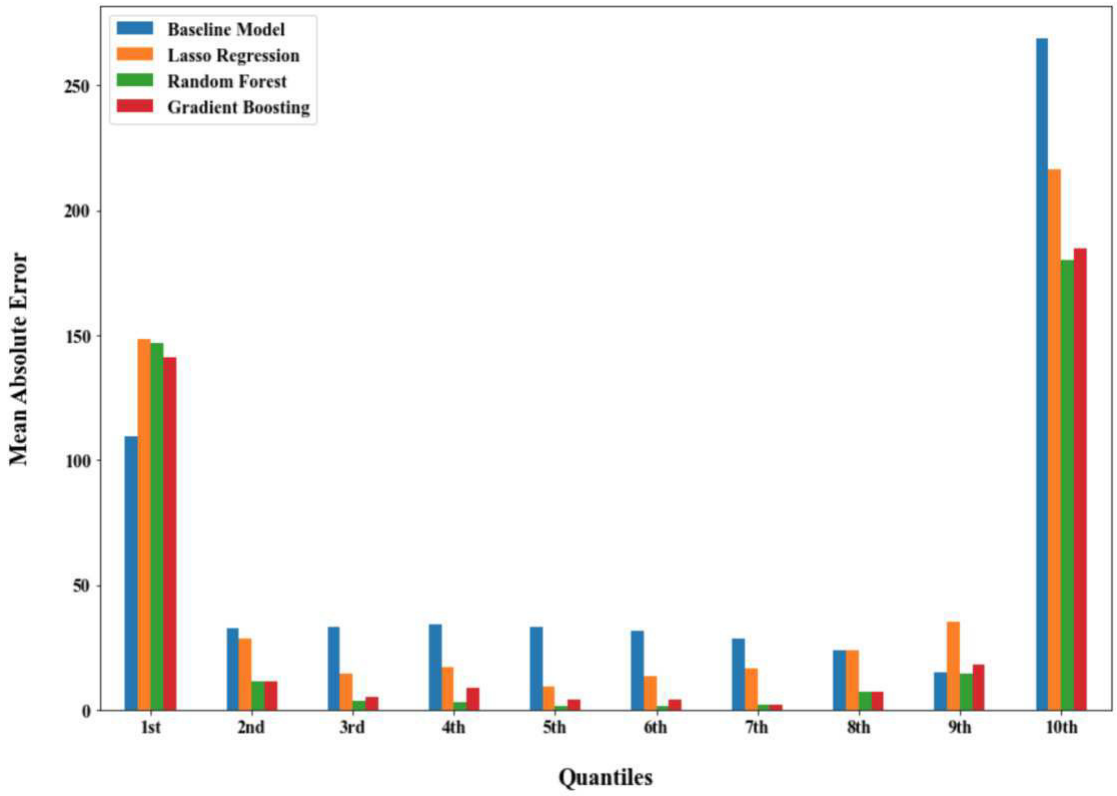
To investigate whether this prediction is accurate for firms that experience a dramatic change in fortune (by transitioning from a low to a high or a high to a low level of competitiveness), we divided the dataset into 10 quantiles based on how much the turnover decreased or increased from 2011 to 2016. This split implied that the 1st quantile contained the 10% of the firms that experienced the greatest decline in turnover and the 10th quantile contained the 10% of firms that experienced the greatest increases in turnover. Table III shows how the intervals of the quantile ranges.

Table III: Mean Absolute Error per Quantile for Four Different Statistical Models

	Quantile Interval	Baseline Model	Lasso Regression	Random Forest	Gradient Boosting
1st	[-15717, -10.63)	109.81	147.18	148.78	145.57
2nd	[-10.63, -2.04)	32.94	28.88	12.22	11.74
3rd	[-2.04, -0.57)	33.54	15.07	3.79	5.1
4th	[-0.57, 0.0)	34.23	17.78	3.21	9.1
5th	[0.0, 0.41)	33.43	9.22	1.48	3.99
6th	[0.407, 1.49)	31.95	13.35	1.81	3.91
7th	[1.49, 3.95)	28.59	16.46	2.4	1.95
8th	[3.95, 10.44)	23.8	24.22	7.39	7.37
9th	[10.44, 31.25)	15.39	36.4	13.93	17.97
10th	[31.25, 8024.86)	268.61	214.03	180.25	182.89

As we can see from Figure 1, we see that all the statistical models (measured by MAE) perform considerably worse for the 1st and 10th quantile than for the other quantiles. In addition, we can see that the machine learning models perform worse than the baseline model in the 1st quantile, and for the 10th quantile, the machine learning models perform only slightly better than the baseline. Based on this performance, we can conclude the machine learning models are not able to provide anywhere near as accurate predictions for firms that experience a dramatic change in competitiveness as for those that experience modest changes.

Figure 1: Comparison of Mean Absolute Error per Quantile for Four Different Statistical Models



5. Interpretations and Conclusions

This article has tried to answer (i) whether information about the stock and flow of employee skills can be used to predict a firm's future competitiveness and (ii) whether these predictions are accurate for firms that experience a dramatic change in fortune.

In terms of predicting a firm's future competitiveness, the analysis found that skills data can indeed be used to improve forecasts of firm competitiveness. We saw that the Lasso, Random Forest and Gradient Boosting models performed significantly better than our baseline model. Nevertheless, the analysis also showed that the naïve econometric model struggled to make sense of the skills data. Only the machine learning models were able to improve the accuracy of our predictions, which illustrates the complexity involved in making use of sparse and high-dimensional skills data.

In terms of identifying whether these predictions are accurate for firms that experience a dramatic change in fortune, the analysis found that the predictive models became less reliable at the "tail ends", where we could find the firms with the highest increase or decrease in turnover. The reason for this decline in reliability is uncertain. It might be that there are some influential observations in the dataset, which distort the analysis, which the large ranges for the 1st and 10th quantiles might indicate (we are indebt the one of the reviewers for pointing out this explanation). It might also be that skills play a less important role in determining extreme reversals of fortune. Other factors, such as market shocks, might be more important for predicting radical changes in competitiveness.

References

Bhawsar, P., and Chattopadhyay, U. (2015), "Competitiveness: Review, Reflections and Directions", *Global Business Review*, **16**(4), 665-679.

Boschma, R., Eriksson, R., and Lindgren, U. (2008), "How does labour mobility affect the performance of plants? The importance of relatedness and geographical proximity", *Journal of Economic Geography*, **9**, 169–190.

Brunello, G., and Wruuck, P. (2019), "Skill Shortages and Skill Mismatch in Europe: A Review of the Literature", IZA Discussion Paper No. 12346, Available at SSRN: https://ssrn.com/abstract=3390340.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, USA: CRC Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning: With applications in R*, New York: Springer.

Kuhn, M., and Johnson, K. (2013), *Applied predictive modeling*, New York: Springer.

Mullainathan, S., and Spiess, J. (2017), "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, **31**(2), 87-106.

Prahalad, C. K., and Hamel, G. (1990), “The core competence of the corporation”, *Harvard Business Review*, **68**(3), 79–91.

Teece, D., Pisano, G., and Shuen, A. (1997), “Dynamic Capabilities and Strategic Management”, *Strategic Management Journal*, **18**(7), 509–533.