A JOURNAL OF THE INSTITUTE FOR OPERATIONS RESEARCH AND THE MANAGEMENT SCIENCES

STRATEGY SCIENCE

Volume 2, Number 1, March 2017

Special Issue Introduction: Evolutionary Logics of Strategy and Organization
William P. Barnett, Daniel A. Levinthal

The Empirics of Learning from Failure
Victor Manuel Bennett, Jason Snyder

How Commitment to Craftsmanship Leads to Unique Value: Steinway & Sons' Differentiation Strategy
Gino Cattani, Roger L. M. Dunbar, Zur Shapira

When More Selection Is Worse
Jerker Denrell, Chengwei Liu, Gaël Le Mens

The Coevolution of Organizational Knowledge and Market Technology
Elizabeth G. Pontikes, William P. Barnett

Strategy Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Empirics of Learning from Failure

Victor Manuel Bennett, Jason Snyder

informs
http://pubsonline.informs.org/journal/stsc/

STRATEGY SCIENCE
Vol. 2, No. 1, March 2017, pp. 1–12
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

The Empirics of Learning from Failure

Victor Manuel Bennett,[a] Jason Snyder[b]

[a] Fuqua School of Business, Duke University, Durham, North Carolina 27708; [b] Eccles School of Business, University of Utah, Salt Lake City, Utah 84112

Contact: vmb10@duke.edu (VMB); jsnyder@eccles.utah.edu (JS)

Received: February 22, 2016
Accepted: September 9, 2016
Published Online: February 20, 2017

https://doi.org/10.1287/stsc.2016.0020

Abstract. The ability to learn from experience is central to an organization's performance. A set of qualitative management studies argues that learning from failure is the exception rather than the rule. Another literature, using econometric methods, finds strongly statistically- and economically-significant effects. There are many possible explanations for this discrepancy, but we argue that one contributor is that a problem with one of the standard empirical approaches to identifying learning from failure may result in erroneously significant results. We generate simulated placebo data in which no learning takes place and show that the standard approach yields strong significant results. We provide a simple example that provides intuition for why this might be. We then propose and implement improved specifications using data on liver transplantation and find no direct evidence of learning from failure.

History: This paper has been accepted for the *Strategy Science* Special Issue on Evolutionary Perspectives on Strategy.

Keywords: strategy • learning • failure • methods

Introduction

Practitioners and scholars have suggested that one significant determinant of an organization's future performance is its ability to learn from prior failures (Argote et al. 1990, Henderson 1974, Pisano et al. 2001). Results on how common it is to successfully learn from failure, however, are mixed.

Literature using archival data largely finds evidence of learning from failure. Studying the satellite launch industry, Madsen and Desai (2010) find that the probability of a successful future launch increases with the cumulative number of failed launches.[1] Both Audia and Goncalo (2007) and Baum and Dahlin (2007) find correlations consistent with cumulative past successes driving organizations' local search for minor performance improvements, and with cumulative past failures leading organizations to search beyond their boundaries for new ideas. Li and Rajagopalan (1997) also find evidence indicating that learning from failure improves quality more than does learning from success.

Results from qualitative fieldwork, on the other hand, suggest that almost any result on learning from failure is surprising. Edmondson (2011) writes:

> When I ask executives to [. . .] estimate how many of the failures in their organizations are truly blameworthy, their answers are usually in single digits—perhaps 2% to 5%. But when I ask how many are treated as blameworthy, they say (after a pause or a laugh) 70% to 90%. The unfortunate consequence is that many failures go unreported and their lessons are lost.

Edmondson suggests that organizations that learn from failure are the exception rather than the norm. Furthermore, Tucker and Edmondson (2003, p. 56), write:

> We conducted a detailed study of hospital nursing care processes to investigate conditions under which nurses might respond to failures they encounter in their hospital's operational processes by actively seeking to prevent future occurrences of similar failures. Our research suggests that, in spite of increased emphasis on these issues, hospitals are not learning from the daily problems and errors encountered by their workers. We also find that process failures are not rare but rather are an integral part of working on the front lines of health care delivery.

This suggests a tension between the archival literature's findings that failure is an important source of future performance improvement and the qualitative literature's suggestion that it is only the rare organization that learns from failures.

There are many possible reasons for the disagreement about likelihood of learning. We suggest that one reason is that previously unknown issues with a standard empirical specification—regressing probability of success on both the cumulative number of successes and the cumulative number of failures to date—can erroneously yield evidence of learning where none exists.

Seeking to resolve the tension, we start by examining how the common empirical specifications in the literature on learning from failure perform using randomly generated data, which, by definition, should show absolutely no effects of learning.[2] For these placebo data sets, we find that using the common empirical specifications in the literature systematically forces strongly significant evidence of learning from failure even though no such learning exists.

We describe two issues with the standard specification that are responsible for the erroneous results. The first is a mechanical relationship that can bias the coefficient, forcing it to deviate from zero on even random data where it should be zero. The second is a problem known in the econometrics literature as the "unit root problem," which can bias significance tests towards suggesting significance too often. While it is beyond the scope of our paper to replicate each study individually, our simulation results suggest caution when interpreting the prior literature's associations.

We go on to (a) suggest an alternate specification for studying learning from failure, and (b) demonstrate this in the context of liver transplantation.

Our approach investigates the effect of a surgery outcome on the next few surgeries, rather than using cumulative counts.[3] We show that this specification correctly yields no result on placebo data with no learning and the correct result on simulated data with learning. We then illustrate the technique on data from liver transplantation surgeries and find little evidence of any learning from failure. In fact, we find that, in some specifications, a prior surgical failure is positively correlated with the chance of a future failure. This stands in stark contrast to the findings of much of the previous literature.

In the next section, we describe the problems with the standard specification. In the third section, we describe our proposed tests of learning from failure and implement them using data on liver transplant surgeries.

Standard Specification for Studying Learning from Failure

Many of the studies attempting to decompose learning-by-doing into learning-from-failure and learning-from-success have used similar empirical models. The standard practice is to model $y_{i,t}$, which is the likelihood of failure of a certain practice at time t for a given organization or individual i, as a function of the cumulative number of successes to date, the cumulative number of failures to date, a set of controls, and a random-error term $e_{i,t}$:

$$y_{i,t} = \alpha + \beta \sum_{s=1}^{t-1} y_{i,s} + \gamma \sum_{s=1}^{t-1} (1 - y_{i,s}) + \eta X_{i,t} + e_{i,t}. \quad (1)$$

Some studies include subtle variations with some of the same mathematical properties. For example, instead of including the count of failures to date, some studies include the sum of losses due to failures, which would be very highly correlated with the count of failures. Others include years of operating experience—which is very highly correlated to the linear combination of cumulative successes and failures—instead of using a count of cumulative successes or failures. Table 1 lists a few example papers, from a selection of strategy journals, using similar specifications.

In the specification from Equation (1), β is to be interpreted as the marginal increase in the likelihood of success at time t as a function of an additional success before time $t-1$. Similarly, γ is interpreted as the marginal increase in the likelihood of success at time t due to an additional failure before time $t-1$. If $\beta \neq \gamma$, that suggests that practitioners learn differentially more from successes than from failures, or vice versa. For example, Staats and Gino (2013) suggest that practitioners might learn more from their own successes than from their own failures, in which case $\beta > \gamma$.

The logic of the existing models is compelling, but it is unclear whether the results of existing empirical analyses are due to real learning from failure in the underlying data or from issues induced mechanically in the standard model.

The standard model has two issues that could yield the observed results even if the relationships were not present in the underlying data. The first problem,

Table 1. A Sample of the Prior Literature

Paper	Journal	Setting	Variables included	Dependent variable
Baum and Dahlin (2007)	*Organization Science*	Railroads	Logged operating experience and logged accident experience	Accident cost per operating mile
Madsen and Desai (2010)	*Academy of Management Journal*	Satellites	Cumulative successes and cumulative failures to date	Launch failure [0,1]
Staats and Gino (2013)	*Management Science*	Heart surgeons	Cumulative successes and cumulative failures to date	Surgery outcome [0,1]
Haunschild and Sullivan (2002)	*Administrative Science Quarterly*	Commerical airlines	Prior accidents and airline age	Number of accidents and incidents per 100 k departures

which to the best of our knowledge has not been explored, is a mechanical relationship that can all but guarantee that the estimated effect of prior failures on likelihood of failure is positive. The second problem is known in the econometrics literature as the "unit root problem" (Hamilton 1994). Algebraically, estimating the standard model is equivalent to estimating a model with a random walk as one of the independent variables. When an independent variable is a random walk, t ratios will no longer be distributed t under the null hypothesis. In practice, what happens is that even with no learning, the t statistics on prior successes and failures will not be mean zero over a large number of random populations.

Next, we describe the two problems and then estimate the standard model on randomly generated placebo data with no learning—including learning from failure—by definition. We find that the two problems lead us to results with strength comparable to that found in the literature.

Simulated Data with No Learning

We begin by generating a placebo data set, which, by definition, has no learning. Figure 1 presents an example of data from the general process we used.[4] We begin by generating 1,000 observations y_t, which are equal to 1 if there is a failure and 0 otherwise. We set the failure probability at 1% to parallel our results from the liver transplant market. This means that for any given random data set, there are, on average, 990 successes and 10 failures. In Figure 1, there are six failures. For each observation, we calculate the cumulative sum of successes and failures to date.[5] The top panel of Figure 1 depicts all 1,000 observations of one simulated data set plotted. The bottom panel shows the first 20 observations in a table, to show how the data is constructed. For any time t, there are three variables: *Cumulative prior failures*, which is the sum of all failures through $t-1$; *Cumulative prior successes*, which is the sum of all successes through $t-1$; and *Failure*, which equals 1 if the event is a failure and 0 otherwise.

The Induced Slope Effect

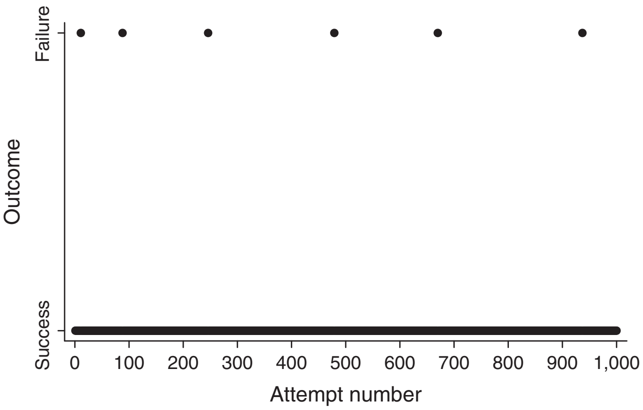
Consider the standard model:

$$\begin{aligned} \textit{Failure}_t = {} & \alpha + \beta_1 \cdot (\textit{Cumulative number of successes})_{t-1} \\ & + \beta_2 \cdot (\textit{Cumulative number of failures})_{t-1} \\ & + e_t. \end{aligned} \qquad (2)$$

Although it is in no way obvious at first glance, the structure of this standard model mechanically induces results, meaning that a deterministic linear relationship between the left and right sides of the equation constrain the coefficients' signs, regardless of the underlying data. We illustrate this using the simulated data from the *Simulated Data with No Learning* section.

To begin, recall that the placebo data are constructed such that there should be no correlation between cumulative experience and probability of failure. Column 1 of Table 2 confirms this is the case.

Figure 1. Example of One Realization of Placebo Data



Coding of variables for first 20 observations from Figure 1

Attempt	Failure	Cumulative prior attempts	Cumulative prior sucessess	Cumulative prior failures	Prior attempt failure
1	0	0	0	0	0
2	0	1	1	0	0
3	0	2	2	0	0
4	0	3	3	0	0
5	0	4	4	0	0
6	0	5	5	0	0
7	0	6	6	0	0
8	0	7	7	0	0
9	0	8	8	0	0
10	0	9	9	0	0
11	1	10	10	0	0
12	0	11	10	1	1
13	0	12	11	1	0
14	0	13	12	1	0
15	0	14	13	1	0
16	0	15	14	1	0
17	0	16	15	1	0
18	0	17	16	1	0
19	0	18	17	1	0
20	0	19	18	1	0

Despite the data being constructed to represent no actual learning, column 2 of Table 2 shows strong significant results on *Cumulative number of successes* and *Cumulative number of failures*.

To clarify the intuition, we describe an analogous model and then show how it is related. Instead of modeling the effect of cumulative prior failures on probability of a current failure, consider a nonparametric model with (a) a dummy for each of the possible numbers of failures to date, and (b) the interactions between those dummies and the number of failures:

$$\begin{aligned} \textit{Failure}_t = {} & \alpha + \beta_1 \cdot (\textit{Cumulative successes})_{t-1} \\ & + \beta_{2,1} \cdot (\textit{One cumulative failure})_{t-1} \\ & + \beta_{2,2} \cdot (\textit{Two cumulative failures})_{t-1} + \cdots \\ & + \beta_{3,1} \cdot (\textit{Cumulative successes}) \\ & \times (\textit{One cumulative failure})_{t-1} + \cdots + e_t. \end{aligned} \qquad (3)$$

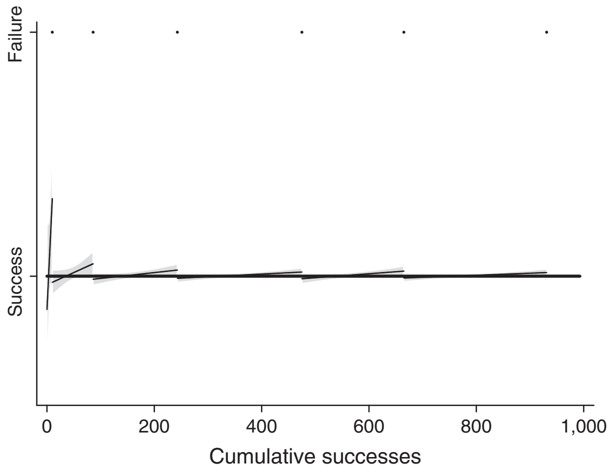
Table 2. Regression Results Run on Placebo Data from Figure 1

	Dependent variable: Failure = 1		
Independent variable	(1)	(2)	(3)
Intercept	0.009399 (0.006106)	0.045401 (0.020455)**	0.090264 (0.08691)
Cumulative number of successes	0.000007 (0.00001)	0.00014 (0.00006)**	0.00013 (0.00006)**
Cumulative number of failures		−0.03107 (0.01367)**	
Cumulative number of failures = 1			−0.08347 (0.08794)
Cumulative number of failures = 2			−0.10516 (0.08766)
Cumulative number of failures = 3			−0.13228 (0.08950)
Cumulative number of failures = 4			−0.15856 (0.09330)*
Cumulative number of failures = 5			−0.18946 (0.09896)*
Cumulative number of failures = 6			−0.21436 (0.10518)**
Cumulative successes dummy variables	No	No	No
R^2	0.0006	0.0197	0.0247
Observations	1,000	1,000	1,000

Note. Parentheses contain robust standard errors using simulated data from Figure 1.

* and **indicate significance at the 10% and 5% confidence levels, respectively.

Figure 2. Intuition for the Induced Slope Effect



Notes. Each line represents a separate regression of failure on cumulative successes within each cumulative failure bin. For example, up to and including the first failure, cumulative failure = 0. After the first failure and up to and including the second failure, cumulative failure = 1. Hence that is a separate bin. The slopes of all of these lines will, by definition, be positive except for the last line, which will have a slope equal to zero. The coefficient on cumulative success in column 3 of Table 3 is the weighted average of each of these slopes.

Equation (3) makes the relationship of this model to a standard fixed-effects model much clearer. Here, $\beta_{3,1}$ can be interpreted as the slope of the regression line estimating the linear effect of successes on only those observations for which there had been only one prior failure.

Figure 2 shows these regression lines on different "bins" of the data, where each bin is a set of observations that share the same number of prior cumulative failures. Note that, by definition, in each bin, every point until the last is a success and the last is a failure. No more than one failure can be included or all of the points would not share the same number of prior failures and thus would not share a slope. The last point must be a failure because otherwise the points after it would have the same number of prior failures as those pictured. In fact, by definition, every bin's regression line except for the last must look the same: n successes followed by one failure. The last bin will not have a failure in it unless the last observation was a failure. As one can see, a fit line in all but the last of those bins will slope upwards. The fit line in the last bin, because all points are successes, would have a slope of zero.

Column 3 of Table 2 shows the regression results from the prior thought experiment.

To move from this analogue to the standard specification, we need to recall the relationship between an aggregate linear effect over a set of subgroups and the linear effects within the subgroups. Algebraically, the slope of the line on the full sample must be a weighted linear combination of the slopes of the lines on the subset. In other words, the aggregate slope will be the sum of each of the slopes of the sum lines multiplied by the percentage of the observations that fall in its bin. A linear combination of a single zero and a set of positive numbers must be positive. This means that regardless of the process that generated these data, the standard specification will always yield a mechanically induced coefficient.

We have demonstrated the induced slope problem in a simplified setting. The addition of other controls or transformations of the data may weaken or break this relationship, but demonstration of the problem urges caution in the interpretation of the models that share this general structure.

The Unit Root Problem

A second potential limitation of the standard specification is that in some settings, it may suffer from a long-studied phenomenon known in time-series econometrics as the "unit root problem" (Dickey and Fuller 1979, Hamilton 1994). The fundamental underlying assumption in using t-tests on regression coefficients is that, under the null hypothesis, the test statistic converges in probability to some known distribution. In the case of studying the effect of cumulative success and cumulative failure on future performance, that means that if many studies were conducted on whether experience affects success—all using the same empirical specification—the regression coefficients from the studies would vary due to randomness. If the null hypothesis were true and experience had no effect on chance of success, the assumptions of ordinary least squares (OLS) would dictate that the t-statistics computed from those statistics would be distributed t with the appropriate number of degrees of freedom. It is necessary to know the distribution to which those t-statistics would converge under the null to know how unlikely it would be to get coefficients as extreme as those observed if the null were true.

The unit root problem is that when one of the independent variables has a unit root (for example, if it has a random walk), t-statistics do not converge to the expected distribution. When estimating models using time-series data in which prior outcomes affect current outcomes, the variance of current outcomes is, in some part, due to the variance in prior outcomes. As long as the effect of those prior outcomes dies out eventually, which happens when the coefficient on a lagged variable is less than 1, the variance of coefficients is stationary. Stationary variance is necessary for t-tests to make sense. If those effects do not die out (meaning that the coefficient on the lagged value is unity), then the variance of the coefficient goes to infinity as T increases. This, in turn, means that the distribution of t-statistics would not converge to what we expect. This means that t values greater than 1.96 occur more than 5% of the time. In fact, in our placebo data, they occur closer to half the time.

The Unit Root Problem and the Standard Specification. To see that the standard specification suffers from the unit root problem, begin with Equation (1), the assumed true data-generating process for a success y_t in the current period:

$$y_t = \alpha + \beta \sum_{s=1}^{t-1} y_s + \gamma \sum_{s=1}^{t-1}(1 - y_s) + \eta X_t + e_t.$$

Then, demean both sums:

$$\begin{aligned} &= \alpha + \beta\left[\sum_{s=1}^{t-1}(y_s - \bar{y}) + \sum_{s=1}^{t-1}\bar{y}\right] \\ &\quad + \gamma\left[\sum_{s=1}^{t-1}(1 - y_s - (1-\bar{y})) + \sum_{s=1}^{t-1}(1-\bar{y})\right] + \eta X_t + e_t; \\ &= \alpha + \beta\left[\sum_{s=1}^{t-1}(y_s - \bar{y}) + (t-1)\bar{y}\right] \\ &\quad + \gamma\left[\sum_{s=1}^{t-1}(1 - y_s - (1-\bar{y})) + (t-1)(1-\bar{y})\right] + \eta X_t + e_t. \end{aligned}$$

Now, define

$$z_t \equiv \sum_{s=1}^{t-1}(y_s - \bar{y}) = \sum_{s=1}^{t-1} e_t$$

and substitute back into Equation (1):

$$y_t = \alpha + (t-1)[\beta\bar{y} + \gamma(1-\bar{y})] + (\beta - \gamma)z_t + \eta X_t + e_t. \quad (4)$$

From Equation (4), we can see a potential challenge with this specification. Note that z_t is a random walk and that, as described in Hamilton (1994, Section 17.1), $\beta - \gamma$ will not converge in distribution to a mean zero normal. That also implies that neither β nor γ will converge in distribution to a mean zero normal and that standard t-tests for $\hat{\beta}$ or $\hat{\gamma}$ being different from zero will not be appropriate. Most importantly, the conventional critical values for judging statistical significance will be wrong.

The unit root problem is less intuitive than the induced slope effect, but the authors have found the following intuition illuminating. Standard econometric testing depends on creating estimators whose distributions, at least in the limit, are known. For example, if we want to test the significance of a regression coefficient $\hat{\beta}$, we convert it into the t-statistic $t = \hat{\beta}/SE$. With enough observations, it can be shown that the t, has a distribution with 95% of its mass between roughly 1.96 and −1.96, under the null. That means that t values outside of that range are unlikely under the null—specifically they occur less than 5% of the time. Thus, central to econometric testing is the assumption that we know to what distribution the statistic converges.

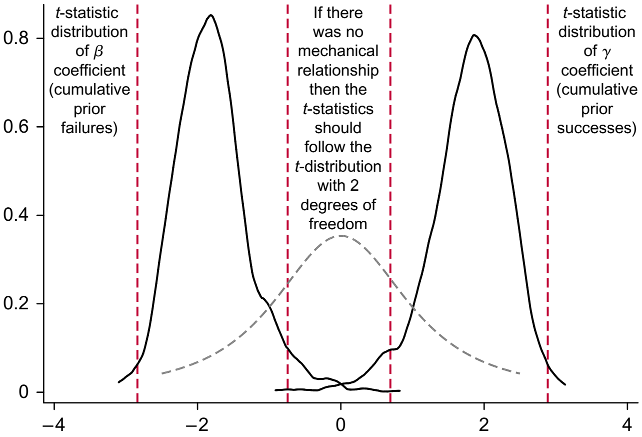
With a unit root process, each observation contains the sum of its own ε and all the ones before it. Even if the error terms ε_i are i.i.d., summing more of them gives larger and larger variances, and thus the t statistic does not converge to the t-distribution. The closest analogue we could construct is determining whether water is boiling by checking its temperature in Fahrenheit, against a Celsius chart. One could be correct, but there is a large range of values for which one would not be correct. We direct researchers who would like more intuition about unit root processes and co-integration to Murray (1994). Researchers whose primary concern is not the statistics, but detecting whether their data series may yield erroneous results, can apply the Dickey-Fuller test (Dickey and Fuller 1979) (clearly described in Hamilton 1994) to determine whether their data series has a unit root, and thus whether a first difference or sliding window specification should be used instead of the standard specification.

Demonstrating the Unit Root Problem on Simulated Data. To demonstrate the unit root problem in practice, we return to Table 2, where we estimated Equation (2) (reprinted here) on the placebo data:

$$\begin{aligned} \mathit{Failure}_t = {} & \alpha + \beta_1 \cdot (\mathit{Cumulative\ number\ of\ successes})_{t-1} \\ & + \beta_2 \cdot (\mathit{Cumulative\ number\ of\ failures})_{t-1} \\ & + e_t. \end{aligned} \quad (2)$$

The way we have constructed the data, $\alpha = 0.01$ and $\beta_1 = \beta_2 = 0$. Note that in column 1, where only cumulative successes are included, the estimated coefficient is—correctly—a very precisely estimated zero. In column 2, using the standard specification, the coefficients are, as expected from the algebra in *The Unit Root Problem and the Standard Specification* section, of opposite sign.[6] Note also that they are indicated as statistically significant because the t-ratios are greater than 1.96, despite the data being constructed in such a way that these variables should not predict success. As described in the previous subsection, this is because the critical values used in conventional t-tests assume that, under the null, the t-ratio will be distributed Student's t. Dickey and Fuller (1979), however, show that the t-ratios from data-generating processes with unit roots are not distributed Student's t. Because the null hypothesis of no correlation is true in our simulated data, we can empirically illustrate this. To do so, we perform 1,000 iterations of the data-generation and regression described above and plot the t-ratios in Figure 3. Note that the bulk of the distribution occurs outside the conventional critical values. In fact, we can see that a finding of zero would occur less than 1% of the time. This means that with this specification, using the conventional critical values will result in erroneously interpreting coefficients as significant in many cases in which they are not.[7]

Figure 3. (Color online) Distribution of t-Statistics on Failure and Success Coefficients from 1,000 Randomly Drawn Data Sets



Notes. t-statistics for robust standard errors at the 95% confidence interval shown. Dashed lines are ±1.96 standard deviations from the mean. The number of observations in each data set equals 1,000. These results follow from the specification:

$$\begin{aligned}(\text{Die within a day of transplant})_t \;=\;& \alpha + \beta_1 \cdot (\text{Cumulative successes})_t \\ &+ \beta_2 \cdot (\text{Cumulative failures})_t + \epsilon_t\end{aligned}$$

This closely follows column 2 of Table 2. The baseline probability of failure is 1%, but this assumption is relaxed with little change in Figure A.1 in the appendix. Figure A.2 increases the size of each random data set; the size of the t-statistic does not change.

The unit root and induced slope problems combined imply that estimating this model on arbitrary data will (a) yield positive coefficients on cumulative successes, and (b) suggest that those effects are significant more liberally than expected. This does not imply that learning from failure does not exist, nor does it necessarily refute studies that have found evidence of it. Those studies all vary from the standard model in various ways, and the transformations of variables or sets of controls used may mitigate these two problems. But our results on simulated data do serve as a caution not to use specifications similar to the standard model and to take care in interpreting results from studies that do.

In the subsequent sections, we suggest alternate methods for testing hypotheses about learning from failure and apply them to data on the success and failure of liver transplant operations.

Demonstrating an Alternate Specification on Simulated Data

The standard specification for studying learning from failure yields biased results and erroneous significance. In its place, we recommend a specification including a sliding window of prior attempts rather than the entire cumulative history, and a count of only failures or success, and not both.

$$Failure_{i,t} = \alpha + \beta_1 \cdot \sum_{x=1}^{n} Failure_{i,t-x} + \eta X_{i,t} + e_{i,t}.$$

This approach mimics the first-differencing advocated by the time-series literature (Hamilton 1994). The constant length of history from using a sliding window prevents the unit root issue. Using only failure or success prevents induced slope problems. Figures 4 and 5 demonstrate the properties of this specification on placebo data with no learning. To generate Figures 4 and 5, we generate 1,000 independent samples of 1,000 random outcomes each, with a failure probability of 10%. On each sample, we estimate our recommended specification with a sliding window of five trials. Figure 4 plots the kernel density of the t-statistics on β_1, which are much closer to the t distribution than with the standard specification (Figure 3). Importantly, the distribution of the t-statistics has its mean at zero, as it should. Figure 5 shows the distribution of coefficients β_1 themselves. As expected, the mean value of the coefficient is zero, as opposed to the coefficients under the induced slope effect (Table 2, column 2).

Figures 4 and 5 demonstrate that our recommended specification does not generate the "false positives" generated by the standard specification. Figure 6 and 7 demonstrate that the specification generates "true positives." For Figures 6 and 7, we generate 1,000 independent samples of 1,000 random outcomes each, with a failure probability of 10% minus 1% per failure in the last five outcomes. This represents short-lived

Figure 4. Distribution of t-Statistics on Simulated Data With No Learning (Placebo)

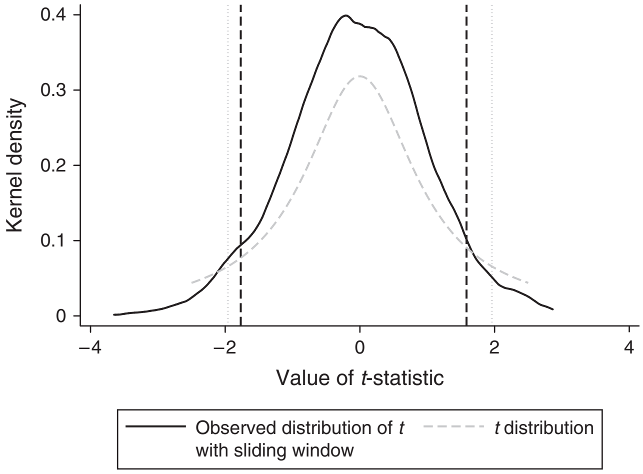


Figure 5. (Color online) Distribution of Coefficient Estimates on Simulated Data With No Learning (Placebo)

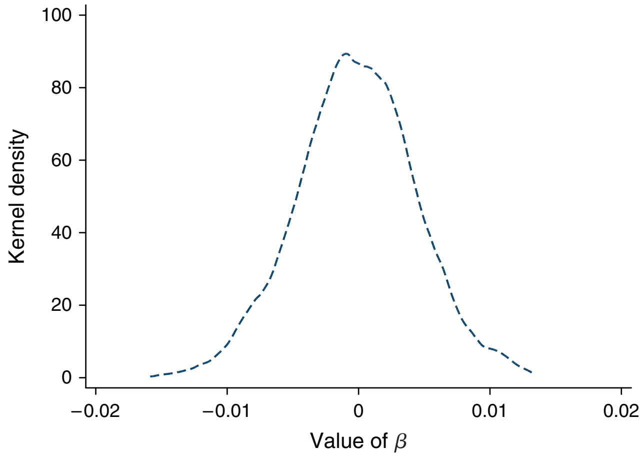


Figure 6. Distribution of t-Statistics on Simulated Data with Learning

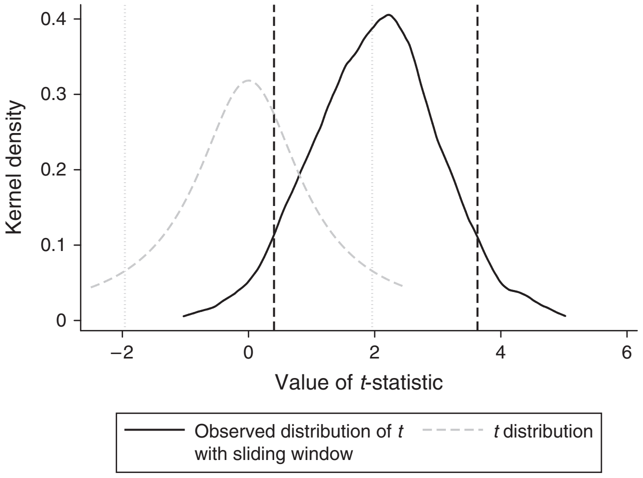
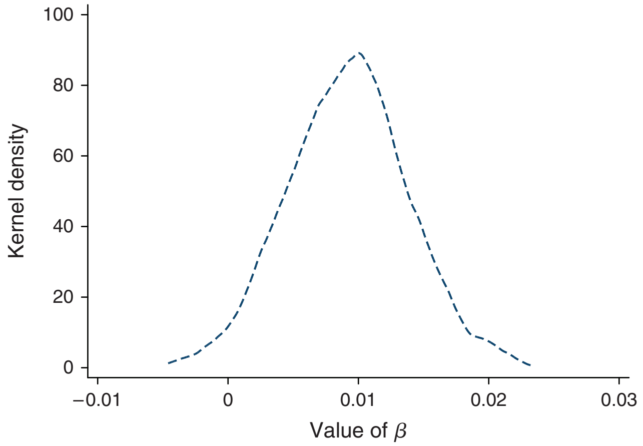


Figure 7. (Color online) Distribution of t-Statistics on Simulated Data with No Learning (Placebo)



learning from failures. Figure 6 depicts the distribution of t-statistics from our recommended specification. The mean of the distribution of t-statistics is greater than 1.96, meaning the coefficient is estimated to be different from zero at greater than 95% confidence. Figure 7 depicts the distribution of β_1. The estimated learning coefficient is, on average, the true value of 1%.

Having demonstrated the properties of our recommended specification on simulated data, we move on to illustrating the specification using real world data from surgical success in liver transplants.

Learning from Failure in Liver Transplants

Data Description

The United Network for Organ Sharing (UNOS) database contains information on every liver transplant performed in the United States since 1987. Uniquely, it provides the exact date of transplant and date of patient death. This allows us to calculate exact survival rates and to control for a host of germane patient and donor characteristics. Table 3 provides summary statistics for survival rates and patient characteristics.

Since we know the day of any given transplant, we know whether transplants that occurred the day immediately prior at the same center were successful. Suppose that, at a given center, a transplant occurred on May 13, 2004 and the patient died within a day. If the next transplant occurred on May 27, 2004, that May 13 surgery would be recorded as "Die within a day of transplant."[8] Since these procedures can take up to 12 hours and we do not know the exact time of the surgery, we code *Die within a day of transplant* as 1 if the social security master file death date for that patient is equal to the day of transplant or the following day.

While ideally we would have physician-level data, the records are only available at the center level.

Replicating Prior Results

In Table 4, we present the results from estimating the standard specification on (a) placebo data generated to replicate the setting and (b) the real UNOS data.

Table 3. Summary Statistics for All U.S. Liver Transplants 1989–2011

Variable	Observations	Mean	Standard deviation	Min	Max
Die within a day of transplant	111,719	0.0109	0.1037	0	1
Die within a week of transplant	111,719	0.0180	0.1331	0	1
Die within a month of transplant	111,719	0.0396	0.1949	0	1
Die within a year of transplant	111,719	0.1130	0.3166	0	1
Days until next transplant	111,690	8.7544	39.4158	1	6,967
Life support at transplant	111,719	0.1071	0.3093	0	1
Re-transplant	111,719	0.1025	0.3033	0	1
Living donor	111,719	0.0406	0.1973	0	1
Age	111,719	45.9371	17.9978	0	84
Age of donor	111,699	35.7637	18.5109	0	92
HCV	111,719	0.2686	0.4432	0	1
Alcoholic cirrhosis	111,719	0.1160	0.3203	0	1
HCC	111,719	0.0923	0.2894	0	1
White	111,719	0.7319	0.4430	0	1
African-American	111,719	0.0929	0.2902	0	1
Hispanic	111,719	0.1229	0.3284	0	1
Female	111,719	0.3790	0.4851	0	1

Notes. Data comes from the United Network for Organ Sharing. Hepatitis C virus (HCV) indicates primary cause of liver failure is hepatitis C. Hepatocellular carcinoma (HCC) indicates that primary cause of liver failure is malignancy. The large maximum value on days until next transplant is driven by centers that close. In our analysis we take logs of this variable, which limits the impact of any one extreme value.

Columns 1 and 2 present the results where the dependent variable is randomly generated for each of the patients with a mean of 1.18%. Thus, by construction, simulated patient deaths are uncorrelated with patient or center characteristics. We iterate this randomization 1,000 times and report the average coefficients. These results are consistent with our prior simulations: on randomly generated placebo data, we find economically and statistically significant evidence suggesting that cumulative past failures lead to future performance improvement and that cumulative past successes are associated with future failures.

In columns 3 and 4, we run the regressions using the real data from the UNOS transplant database. We find that the parameters are approximately half the size of the parameters of the simulated results, but still highly significant. A naïve reading of these columns would suggest that the real data shows important correlations consistent with the prior literature. Knowing about the induced slope and unit root problems, however, we cannot tell whether these results are evidence of true learning from failure or symptoms of mathematical issues.

Demonstrating Our Recommended Specification on Real Data

Table 5 shows the results of our first suggested empirical approach to solving the learning-from-failure

Table 4. Results on Placebo vs. Actual Transplant Data

	Dependent variable: Die within a day of transplant			
	Placebo data		Real data	
Independent variable	(1)	(2)	(3)	(4)
Cumulative single day deaths	−0.3488	−0.3493	−0.1726	−0.1764
(in 100s of patients)	(0.0620)***	(0.0620)***	(0.0359)***	(0.0370)***
Cumulative single day survival	0.0040	0.0040	0.0018	0.0018
(in 100s of patients)	(0.0008)***	(0.0008)***	(0.0006)***	(0.0006)***
Individual controls		Yes		Yes
Year effects	Yes	Yes	Yes	Yes
Center effects	Yes	Yes	Yes	Yes
Clusters	99	99	99	99
Observations	41,704	41,704	41,704	41,704

Notes. Parentheses contain standard errors clustered at the center level. Data comes from the United Network for Organ Sharing for 1989–2011. Observations restricted to centers that started in 1989 or afterward to avoid left-hand-side censoring for the cumulative variables. Placebo data randomly assigns the dependent variable with probability 1.18%, which is the overall probability that a patient will die within one day for the subsample of centers that started in 1989 or afterward. This process was iterated 1,000 times and the average parameter estimates and standard errors are in the first two columns.

*, **, and *** indicate significance at the 10%, 5%, and 1% confidence levels, respectively.

Table 5. Learning from the Previous Transplant

	Die within a day of transplant		Die within a week	Die within a month	Die within a year
Independent variable	(1)	(2)	(3)	(4)	(5)
Death within a day of transplant for prior surgery	0.0074 (0.0036)**	0.0079 (0.0036)**	0.0107 (0.0046)**	0.0081 (0.0062)	−0.0052 (0.0075)
On life support		0.0153 (0.0018)***	0.0310 (0.0028)***	0.0607 (0.0038)***	0.0932 (0.0057)***
Re-transplant		0.0113 (0.0015)***	0.0187 (0.0022)***	0.0547 (0.0037)***	0.1094 (0.0052)***
Living donor		−0.0069 (0.0011)***	−0.0088 (0.0015)***	−0.0062 (0.0023)***	−0.0041 (0.0050)
Age of patient		0.0002 (0.0000)***	0.0002 (0.0000)***	0.0006 (0.0001)***	0.0018 (0.0001)***
Age of donor		0.0001 (0.0000)***	0.0001 (0.0000)**	0.0002 (0.0000)***	0.0007 (0.0001)***
HCV		−0.0029 (0.0009)***	−0.0034 (0.0011)***	−0.0096 (0.0014)***	0.0106 (0.0035)***
Alcoholic cirrhosis		−0.0026 (0.0011)**	−0.0043 (0.0014)***	−0.0064 (0.0022)***	−0.0105 (0.0033)***
HCC		−0.0038 (0.0012)***	−0.0050 (0.0014)***	−0.0115 (0.0021)***	0.0005 (0.0041)
White		−0.0002 (0.0014)	0.0009 (0.0018)	0.0036 (0.0026)	0.0120 (0.0045)***
African American		0.0004 (0.0015)	0.0010 (0.0023)	0.0021 (0.0036)	0.0168 (0.0050)***
Hispanic		−0.0014 (0.0015)	0.0003 (0.0019)	0.0004 (0.0032)	−0.0014 (0.0045)
Female		0.0011 (0.0006)*	0.0015 (0.0009)*	0.0004 (0.0013)	−0.0025 (0.0021)
Year effects	Yes	Yes	Yes	Yes	Yes
Center effects	Yes	Yes	Yes	Yes	Yes
Clusters	154	154	154	154	154
R^2	0.005	0.009	0.015	0.029	0.042
Observations	111,619	111,599	111,599	111,599	111,599

Notes. Parentheses contain standard errors clustered at the center level. Data comes from the United Network for Organ Sharing for 1989–2011, including all centers in the sample.

*, **, and ***indicate significance at the 10%, 5%, and 1% confidence levels, respectively.

problem: exploiting the high-frequency nature of the liver transplant data set. Instead of using the variable *Cumulative single day deaths*, which measures the total number of such deaths at the center over its history, we use an indicator of whether the prior transplant resulted in the patient dying within a day:[9]

$$(Die\ within\ a\ day\ of\ transplant)_{i,t} = \alpha + \beta_1 \cdot (Death\ within\ a\ day\ of\ transplant)_{i,t-1} + \eta X_{i,t} + e_{i,t}. \qquad (5)$$

In columns 1 and 2 of Table 5, we find that, if anything, a death within a day of the prior transplant patient is positively correlated with whether the current patient dies within a day of transplant. Given the baseline probability of death within a day of transplant of 1.18%, a 0.74% increase in the death probability implies an increased risk of death of approximately 70%. The standard errors are large, however. The probability of immediate death following another immediate death could be anywhere between 1% and 2.5%. We find similar results in column 3, but much less significant results as the measure of mortality becomes broader in columns 4 and 5.

In Figure 8, we expand on specification 5 by calculating coefficients on whether there was a death within a day during each of the 10 previous transplants:

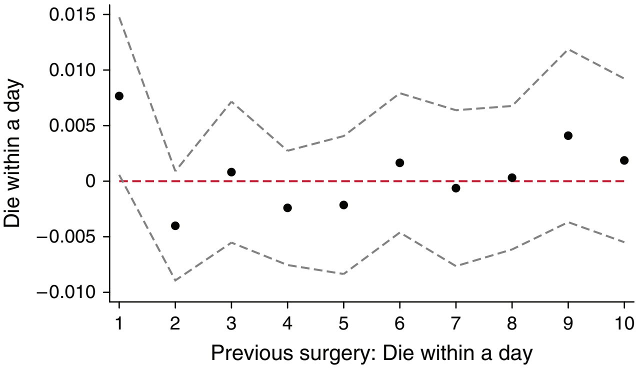
$$(Die\ within\ a\ day\ of\ transplant)_{i,t} = \alpha + \beta_1 \cdot (Death\ within\ a\ day\ of\ transplant)_{i,t-1} + \cdots + \beta_{10} \cdot (Death\ within\ a\ day\ of\ transplant)_{i,t-10} + \eta X_{i,t} + e_{i,t}.$$

Using the specification in column 2 of Table 5, we find that, in general, prior failures other than the failure of the immediately preceding surgery do not have an association with future success. Consistent with the findings of Tucker and Edmondson (2003) and Edmondson (2011), the transplant data presents little evidence of learning from failure.

Discussion

The main goal of this manuscript is to contribute to the strategy research community's attempt to understand

Figure 8. (Color online) Coefficients on Failure in the Previous 10 Transplants



Notes. 95% confidence intervals with standard errors clustered at the center level shown in dashes. These results follow from the specification:

$$\begin{aligned}&(\text{Die within a day of transplant})_{i,t}\\&\quad= \alpha + \beta_1 \cdot [(\text{Previous surgery})_{-1}\text{: (Die within a day)}_i]\\&\qquad + \beta_2 \cdot [(\text{Previous surgery})_{-2}\text{: (Die within a day)}_i] + \cdots\\&\qquad + \beta_{10} \cdot [(\text{Previous surgery})_{-10}\text{: (Die within a day)}_i]\\&\qquad + \gamma \cdot (\text{controls, center, and year effects})_{i,t} + \epsilon_{i,t}\end{aligned}$$

This closely follows column 2 of Table 4, except lagged previous surgeries are included. The findings suggest that the effect of a death in the last surgery may predict a death in the present surgery, but more distant failures are not correlated with current outcomes.

learning from failure. Taking a step back, one main reason for wanting to understand learning from failure is to provide normative recommendations on deliberate strategic actions that firms can take to help individual employees learn from mistakes, or to learn from mistakes at the organizational level. Given that, we propose that researchers studying learning from failure follow a technique long-used in strategy literature to study the performance implications of management practices: to investigate differences in the elasticity of performance to environmental features as a function of management practices. For example, Bennett (2013) investigates the elasticity of prices negotiated by a car dealership to the willingness of the customers to pay as a function of staffing practices. A practice that increases the elasticity of price to willingness to pay is one that increases the proportion of value created that the firm can capture. In a learning setting, this could translate to regressing likelihood of success on count of failures and an interaction with an indicator for organizational practice. The coefficient on the interaction could be interpreted as a measure of whether the practice improves learning from failure. This empirical strategy does not have the aforementioned problems as long as count of successes is not also included. A number of articles have been written on practices organizations can implement to learn from their failures (Desai 2016, Edmondson 2011, Sitkin 1992). A possibly even larger body of literature, however, details the impediments that keep organizations from learning from failure (e.g., Baumard and Starbuck 2005, Haunschild and Sullivan 2002, Kim et al. 2009, Tucker and Edmondson 2003).

Conclusion

We contribute to reconciling archival results suggesting significant organizational learning from failure with qualitative results suggesting that very little such learning occurs. We find that a standard empirical specification used for testing learning from failure has two structural problems that can lead to biased coefficients and biased significance tests. Our results suggest that caution is needed when interpreting the literature's prior results.

The empirical issues with both the tests and the coefficients using the standard specification are certainly not the only possible explanations for the divergence between the econometric and qualitative research. It is entirely possible that differences in the settings or any number of other factors may be involved. These results do not suggest that learning from failure is not happening, but rather that we should interpret results of the standard specification with care.

Suggesting issues with the standard approach to studying learning from failure raises the question of what is the correct way to estimate learning from failure. Unfortunately, there is not an obvious answer. To the best of our knowledge, the issues raised have not been discussed in the econometric literature and the finite sample properties of some of the candidate estimators are not well understood. While we hope future econometric research yields consistent unbiased estimators, we are not left without options for current research.

Our paper shows some initial attempts at bias-free tests of the learning-from-failure hypothesis. We hope that future studies develop these approaches further to provide a more robust understanding of a critical human and organizational phenomenon: learning from failure.

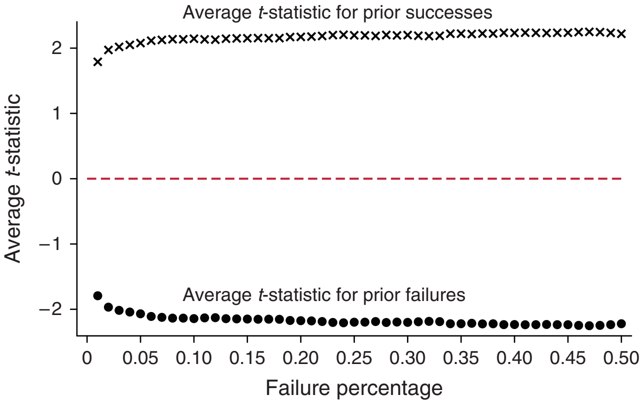
We propose an option for studying learning from failure: using a sliding window of a theory-driven constant length for the regressor so that the error variance is constant.

Acknowledgments

The authors thank Hyungsik Roger Moon for his generosity with his time and comments on empirical issues; Mike Ryall, Connie Helfat, and participants at the Stanford GSB Evolutionary Perspectives in Strategy Conference for comments on prior drafts; and Dr. Scott Palmer, Dr. Margaret Siska, and Ian Jamieson for their time and their knowledge of the organ transplant setting. All mistakes are the authors' own.

Appendix

Figure A.1. (Color online) Average t-Statistic on Failure and Success Parameters from 1,000 Random Data Sets with 50 Different Baseline Failure Rates

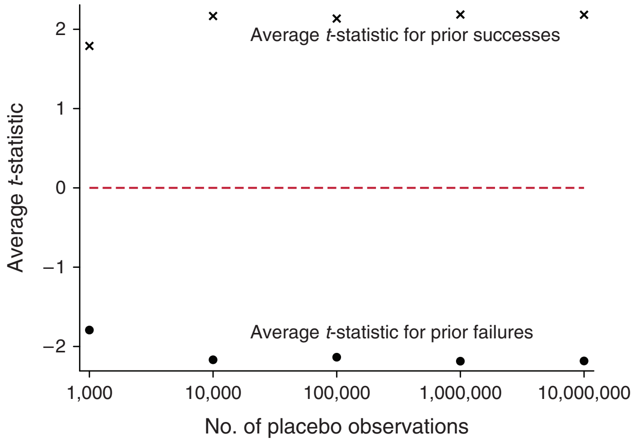


Notes. t-statistics for robust standard errors at the 95% confidence interval shown. These results follow from the specification:

$$\begin{aligned}(\text{Die within a day of transplant})_t &= \alpha + \beta_1 \cdot (\text{Cumulative successes})_t \\ &\quad + \beta_2 \cdot (\text{Cumulative failures})_t + \epsilon_t\end{aligned}$$

This closely follows column 2 of Table 2. The findings suggest that, if anything, larger baseline failure rates increase the average t-statistic on both cumulative failures and cumulative successes.

Figure A.2. (Color online) Average t-Statistic on Failure and Success Parameters from 1,000 Random Data Sets with Different Sample Sizes



Notes. t-statistics for robust standard errors at the 95% confidence interval shown. These results follow from the specification:

$$\begin{aligned}(\text{Die within a day of transplant})_t &= \alpha + \beta_1 \cdot (\text{Cumulative successes})_t \\ &\quad + \beta_2 \cdot (\text{Cumulative failures})_t + \epsilon_t\end{aligned}$$

This closely follows column 2 of Table 2. The findings suggest that the order of magnitude does not determine the size of the t-statistic.

Table A.1. Results with Alternative Specifications on Placebo vs. Actual Transplant Data

	Dependent variable: Die within a day of transplant			
	Placebo data		Real data	
Independent variable	(1)	(2)	(3)	(4)
Cumulative single day deaths (in 100s of patients)	−0.1015 (0.0328)***	−0.3532 (0.0627)***	−0.0932 (0.230)***	−0.1782 (0.0006)***
Age of center	0.0007 (0.0002)***		0.0005 (0.0002)**	
Cumulative total transplants (in 100s of patients)		0.0040 (0.0008)***		0.0018 (0.0007)***
Individual controls	Yes	Yes	Yes	Yes
Year effects	No	Yes	No	Yes
Center effects	Yes	Yes	Yes	Yes
Clusters	99	99	99	99
Observations	41,704	41,704	41,704	41,704

Notes. Parentheses contain standard errors clustered at the center level. Data comes from the United Network for Organ Sharing for 1989–2011. Observations restricted to centers that started in 1989 or later to avoid left-hand-side censoring for the cumulative variables. Placebo data randomly assigns the dependent variable with probability 1.18%, which is the overall probability that a patient will die within one day for the subsample of centers that started in 1989 or later. This process was iterated 1,000 times and the average parameter estimates and standard errors are in the first two columns.

*, **, and *** indicate significance at the 10%, 5%, and 1% confidence levels, respectively.

Endnotes

[1] A notable exception is Staats and Gino (2013), who find almost exactly the opposite at the individual level. They suggest that individual surgeons' future success is positively influenced by their past cumulative number of successes and negatively influenced by their past cumulative number of failures.

[2] We follow the statistics and econometrics literatures and use the term "placebo data" to describe randomly generated data where the effect is absent by design.

[3] This approach is analogous to the first-differencing approach recommended in the time-series literature.

[4] See code for seeds used and procedures for randomization.

[5] The $Z(t)$ statistics from the augmented Dickey-Fuller test for the cumulative success and cumulative failure series are −0.142 and −0.087, respectively. These values are well below the five-percent critical value of −2.86, indicating that we cannot reject the null that the series has a unit root.

[6] In unreported regressions, we estimated a variant of the standard specification in which either cumulative successes or cumulative failures are included—but not both—along with the number of trials or the organization's age as in, for example, Desai (2016), Haunschild and Sullivan (2002), and Ingram and Baum (1997). These results are in appendix Table A.1. The results are qualitatively the same.

[7] We show in Figures A.1 and A.2 that these results on the simulation are robust to alternative baseline models.

[8] If two transplants occurred on the same day, we code "Die within a day of transplant" for the prior surgery as 1 if either patient died within a day of transplant.

[9] Note that counting a surgery as a failure if the patient died within a day is a conservative measure. In Table 5, we show results for alternate definitions of failure, ranging from patients dying within a day to patients dying within a year.

References

Argote L, Beckman SL, Epple D (1990) The persistence and transfer of learning in industrial settings. *Management Sci.* 36(2):140–154.

Audia PG, Goncalo JA (2007) Past success and creativity over time: A study of inventors in the hard disk drive industry. *Management Sci.* 53(1):1–15.

Baum JAC, Dahlin KB (2007) Aspiration performance and railroads' patterns of learning from train wrecks and crashes. *Organ. Sci.* 18(3):368–385.

Baumard P, Starbuck WH (2005) Learning from failures: Why it may not happen. *Long Range Planning* 38(3):281–298.

Bennett VM (2013) Organization and bargaining: Sales process choice at auto dealerships. *Management Sci.* 59(9):2003–2018.

Desai V (2016) Learning to learn from failures: The impact of operating experience on railroad accident responses. *Industrial Corporate Change* 25(2):199–226.

Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* 74(366):427–431.

Edmondson AC (2011) Strategies of learning from failure. *Harvard Bus. Rev.* 89(4):48–55, 137.

Hamilton JD (1994) *Time Series Analysis* (Taylor & Francis, New York).

Haunschild PR, Sullivan BN (2002) Learning from complexity: Effects of prior accidents and incidents on airlines' learning. *Admin. Sci. Quart.* 47(4):609–643.

Henderson B (1974) The experience curve—Reviewed. Accessed January 9, 2015, https://www.bcgperspectives.com/content/Classics/strategy_supply_chain_management_experience_curve_reviewed_the_concept/.

Ingram P, Baum JAC (1997) Opportunity and constraint: Organizations' learning from the operating and competitive experience of industries. *Strategic Management J.* 18(S1):75–98.

Kim J-Y, Kim J-Y, Miner AS (2009) Organizational learning from extreme performance experience: The impact of success and recovery experience. *Organ. Sci.* 20(6):958–978.

Li G, Rajagopalan S (1997) The impact of quality on learning: An empirical study. *J. Oper. Management* 15(3):181–191.

Madsen PM, Desai V (2010) Failing to learn? The effects of failure and success on organizational learning in the global orbital launch vehicle industry. *Acad. Management J.* 53(3):451–476.

Murray MP (1994) A drunk and her dog: An illustration of cointegration and error correction. *Amer. Statistician* 48(1):37–39.

Pisano GP, Bohmer RMJ, Edmondson AC (2001) Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac Surgery. *Management Sci.* 47(6):752–768.

Sitkin S (1992) Learning through failure: The strategy of small losses. *Res. Organ. Behav.* 14:231–266.

Staats BR, Gino F (2013) Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. *Management Sci.* 59(11):2435–2449.

Tucker AL, Edmondson AC (2003) Why hospitals don't learn from failures. *California Management Rev.* 45(2):55–72.

Victor Manuel Bennett is an assistant professor in the strategy area faculty at Duke's Fuqua School of Business. His research focuses on how firms respond to competition including both external actions like price and quality changes, and internal actions like changing staffing, compensation, or the autonomy of staff. His research has been published in top journals including *Management Science*, *Strategic Management Journal*, and *Organizational Behavior and Human Decision Processes*. He previously worked at the University of Southern California, Google, and served as senior economist for the White House Council of Economic Advisers. Professor Bennett earned his bachelor's degree from Stanford University and his doctorate from Berkeley's Haas School of Business.

Jason Snyder is an assistant professor in the Entrepreneurship and Strategy group at University of Utah Eccles School of Management. His research interests include firm strategy, socially responsible business, and political economy. His research focuses on how ethics are transmitted within firms, how competition influences socially responsible behavior, and on how political power is passed along within families. He has published in top journals including *Management Science* and *The Review of Economic Studies*.