COLORADO SCHOOL OF MINES
EARTH • ENERGY • ENVIRONMENT

Division of Economics and Business
Working Paper Series

Tales Left Tails Tell: Addressing Left-Side Truncation in Strategy Research

Richard A. Hunt
Daniel A. Lerner

Working Paper 2017-04
http://econbus-papers.mines.edu/working-papers/wp201704.pdf

Colorado School of Mines
Division of Economics and Business
1500 Illinois Street
Golden, CO 80401

October 2017

Colorado School of Mines
Division of Economics and Business
Working Paper No. **2017-04**
October 2017

Title:
Tales Left Tails Tell: Addressing Left-Side Truncation in Strategy Research*

Author(s):
Richard A. Hunt
Division of Economics and Business
Colorado School of Mines
Golden, CO 80401
rahunt@mines.edu

Daniel A. Lerner
Deusto Business School
Universidad Deusto
Bilbao, Spain

ABSTRACT
This study takes up the issue of left-side truncation (LST) in strategy research. Since reliable data on early-stage and failed firms are often unavailable for analysis, empirical findings are often derived from observations involving survivors. But, what if the missing data bear little resemblance to that which is available? Leveraging a dataset that comprehensively captures an entire industry, including firms that commonly pass unobserved, we demonstrate how LST materially affects empirical results and theoretical acuity. Our analysis suggests that the tools typically used to address LST have flaws that hinder their applicability to many of key questions posed by strategy scholars. In response, and to facilitate more thorough reporting of truncation effects, we present an effective, easy-to-use Truncation Factor.

Keywords: Missing data, Left-side truncation, Research methods, Strategic management, Entrepreneurship, Truncation factor, Non-ignorable data, Statistics

*Hunt is corresponding author.

Tales Left Tails Tell:
Addressing Left-Side Truncation in Strategy Research

ABSTRACT

This study takes up the issue of left-side truncation (LST) in strategy research. Since reliable data on early-stage and failed firms are often unavailable for analysis, empirical findings are often derived from observations involving survivors. But, what if the missing data bear little resemblance to that which is available? Leveraging a dataset that comprehensively captures an entire industry, including firms that commonly pass unobserved, we demonstrate how LST materially affects empirical results and theoretical acuity. Our analysis suggests that the tools typically used to address LST have flaws that hinder their applicability to many of key questions posed by strategy scholars. In response, and to facilitate more thorough reporting of truncation effects, we present an effective, easy-to-use *Truncation Factor*.

KEY WORDS: Missing data, Left-side truncation, Research methods, Strategic management, Entrepreneurship, Truncation factor, Non-ignorable data, Statistics

INTRODUCTION

What if the list of runners who started a long distance race was generated solely from the list of runners who finished the race? Most people would agree that such an approach is riddled with flaws. Consider the case of Siberia's 13.1-mile Baikal Ice Half-Marathon. In 2012, fewer than 10% of the 134 starters ultimately crossed the finish line. It seems safe to imagine that the experiences of the 92% who failed to finish differed, perhaps quite markedly, from the 11 who did. And yet, without information regarding the 123 non-finishers, the average competitor performance would consist solely of data drawn from the 11 finishers. Implicitly, the excluded non-finishers would each be assigned the mean performance of those who actually finished. But, for the 15 runners who did not even make it past the first mile, an imputed assignment of the average for those who finished the entire race is clearly far-fetched. Therefore, failure to take into account the non-finishers would result in a drastic mischaracterization of the average performance and even the nature of the race itself.

What if, instead of a running race, the data under consideration involved a new, high-risk technology, such as early-stage horseless carriage companies pursuing internal combustion engines at the end of the 19th century? Or, what about all the firms that have been involved in cloud computing since the advent of compatible timeshare systems in 1961? Even if one could discern retrospectively all the successful firms and a subset of the failed firms, an incomplete roster of participants makes it difficult to draw reliable conclusions about the phenomenon under investigation. Here too, each unobserved participant is implicitly assigned the mean value of observable participants even though the exact number and characteristics of missing firms are unknown. This is the essence of left-side truncation (LST). Sometimes referred to as "the survivor's bias" (Heckman 1979), LST is a persistent confound for strategy scholars.

Truncation is best understood as a situation in which "observations on both the dependent variable and regressors are lost" (Cameron & Trivedi, 2005). It is essentially a specific instance of the missing data problem (Little & Rubin, 1987), but with a larger scale and fewer remedies. Unlike missing survey items or randomly omitted values, the truncation experienced by strategy scholars often eliminates from consideration research subjects (e.g. firms, founders, or technical innovations) that fail to be included in the population of observable events. Most often, truncation is simply an artifact of circumstances in which the data for successful subjects is observed, while the data for the unsuccessful often evaporates. Ultimately, the adequacy of any dataset is a function of its reliability in addressing the research question being posed. Since LST limits the completeness of the data that is available for analysis, the extent to which missingness constitutes a concern depends on what is being studied. In some instances, the firms that are worth including in a sample are simply comprised of entities that have become operationally active and competitively relevant. Depending upon the specific line of inquiry, organizations that never become substantively operational may be of little consequence.

However, in the realm firm foundings, new sector development, opportunity emergence, and unforeseen technological disruptions, information about both failed and successful attempts are germane. Data on failed attempts are particularly important for scholars studying nascent-stage events, such as: the types of founding groups that form new ventures; the reasons that similar organizations arising within the same environment meet with different fates; the development and implementation of differential entry strategies; or, duration effects related to the time gap between opportunity identification and opportunity development. "Understanding why the fates of startup attempts differ so dramatically has been a major concern of scholars over the past several decades," noted Yang and Aldrich (2012:477), but complete data is often elusive (Amburgey and Rao 1996) even while it is essential (Short, et al., 2010; Ketchen et al. 2008).

This paper reframes and extends prior work on truncation while making two primary contributions. First, using a non-simulated, uniquely comprehensive dataset reflecting new firm entry, operational activity, and survival, our analysis provides a clear illustration of both the empirical and theoretical hazards of left-side truncation. Second, the paper presents a simple *Truncation Factor* robustness test, for use when research might be impacted by non-ignorable missing data. This *Truncation Factor* is easily calculable, and objectively indicates the extent to which the results are susceptible to or secure from the effects of LST.

TRUNCATION IN STRATEGY RESEARCH

By its very nature, strategy research often involves retrospective analyses. Absent the rare occurrence of exogenous events that give rise to complete populations, conclusions are necessarily, and often accurately, drawn from a sub-set of retrievable data. In many of these cases, the absence of complete information may be inconvenient but readily surmountable given a reasonable application of standard parametric assumptions related to normality and distribution

(Roderick & Rubin, 1987). There are, however, important research questions, grounded in retrospective data that are particularly susceptible to the ill effects of data truncation. For example, nascent-stage business activity is one area that is famously beset by missing archival data, and with great consequence (Amburgey & Rao, 1996). An analysis of firm foundings and new market entry that examines primarily surviving firms is apt to yield different findings than a study that includes the few leaders, the trailing participants, and the many missing-in-action (Hannan & Carroll, 1992). Conceptually, LST increases the risk that explanatory frameworks for the development and implementation of strategy and strategic entrepreneurship (Hitt, Gimeno & Hoskisson 1998) are based on only survivors and are thus incomplete.

While left-side truncation poses a potential challenge to all retrospective designs using archival data, the effect on the study of nascent events is particularly problematic (Delacroix & Carroll, 1983; Hannan & Carroll, 1992), including founding teams, new firms and sectors, emergent technologies and even management history (Hunt 2013a, 2013c; Hunt & Ortiz-Hunt 2017). Figure 1 displays a stylized representation of the truncation problem. The individual data points represent a complete population of firms entering a hypothetical market. The truncation threshold represents the beginning of *observability*. Prior studies have established observability through the use of milestones that are indicative of firms becoming substantively operational, such as appearing in trade catalogues (Klepper, 2001) or receiving VC financing (Chatterji, 2009). The issue with building a sample based on such observable milestones, and then working backwards to find the firm origins, is that even the most diligent research will truncate early failures (e.g. entrants failing to become operational or receive VC investment). Since unobserved early-stage firm failures inherently have shorter lifespans than the observable survivors that achieve selected milestones, truncation is not a matter of *if* the average longevity among observable firms is overstated but rather *how significant* the inevitable overstatement is.

[INSERT FIGURE 1 ABOUT HERE]

In Figure 1, the mean performance (Y_O), comprised of only the observed values of y (y_o), overstates the operational performance and understates the entry rate of the true, complete population. This is because of an *unobservable non-linear relationship* in the complete population of observations. Since the unobserved values of y (y_u) do not share a linear relationship with the population of values comprising Y_O, conventional efforts to relate Y_O and Y_U will be compromised by the inability to apply the distributional assumptions underlying parametric correction tools, such as the Tobit Model (Tobin, 1958). Our characterization theorizes that tools to examine nascent-stage organizational phenomena as samples with normally distributed predictors are highly susceptible to misspecification. This is consistent with recent challenges to Gaussian distribution assumptions demonstrated by Crawford et al. (2015). Looking past assumptions of normality, corrections that relax distributional assumptions such as maximum-likelihood estimation and the Heckman two-step estimator are marginally more robust (Heckman, 1979), but suffer from the notable liability of being "fragile to even very minor misspecification of error distributions" (Cameron & Trivedi, 2005). "Even with small amounts of unrecognized nonlinearity," noted Achen, "violently incorrect inferences can occur" (2005; 334). And, since the truncation of actors failing to achieve a milestone is non-random (Kim & Lai, 2000), statistical bootstrapping techniques (Bilker & Wang, 1997) cannot remedy the matter.

The development of semiparametric estimators (Tsiatis, 2006) constitutes yet another avenue to address this dilemma. While parametric models rest upon the core assumption that the error terms are normally distributed, semi-parametric models are not so constrained. Instead, by formulating estimations in the context of less stringent, non-normal distributional assumptions, semiparametric estimators provide improved robustness due to the fact that the models include a combination of both parametric and nonparametric regressors, allowing the model to account for non-linear relationships, including those that may be unobservable. There is, however, a price tag for this analytical flexibility. Specifically, semi-parametric models generate multiple solutions that are comparatively less stable, meaning they are more sensitive to relatively small changes in estimators than is the case when normality can be assumed. Since the boundaries of possible solutions are so wide-ranging as to include the parametric estimators that are insufficiently robust and the non-parametric estimators that are insufficiently realistic or informative, these features render semiparametric estimators problematic to interpret (Tsiatis, 2006) and unsatisfying to apply when conditions of left-side data truncation potentially exist.

For a time-dependent covariate, the Andersen, Borgen, Gil & Keiding, (1993) partial likelihood approach can be used to correct for truncation *if* the entire covariate history is available for all subjects. In strategy research this is seldom the case as some longitudinal covariates are observed intermittently and at discrete time points. Yang and Aldrich (2012) demonstrated truncation's biasing effects and offered corrective guidance for estimating the hazard of early termination. While this approach constitutes an important advance, such corrections are impossible when non-random events (e.g. firm exit) occur prior to the first panel observation. Furthermore, the corrections are still based on information drawn from the population of observed actors, which can be sharply divergent from unobserved events (Cameron & Trivedi, 2005), especially those involving early failures.

LEFT-SIDE TRUNCATION: AN ILLUSTRATION

To illustrate the threats posed by truncation, we offer a detailed example using non-simulated data in the context of industry entry, focusing on entrepreneurial spinoffs. Intra-industry entrepreneurial spinoffs occur when employees leave a (parent) firm to start a new, completely independent company as an entry vehicle into the same industry as their former employer (Klepper, 2001), without support from the parent-firm. Our study contrasts the operational performance and survival of such spinoffs to *de novo* entrants, which are "entrepreneurial companies whose founders have no previous employment ties to other firms in the industry" (Helfat & Lieberman, 2002: 730).

The examination of intra-industry entrepreneurial spinoffs offers a set of conditions that are ideal for a vigorous test of data truncation for several reasons. First, spinoffs are a common occurrence, sometimes representing more than 70% of the total market entrants in certain sectors (Klepper 2001, 2009). Second, by their very nature, entrepreneurial spinoffs involve founding teams, market entry decisions and early failures, all of which are notoriously challenging to systematically observe because documentation is often missing for nascent-stage ventures, especially data related to failures that occur within the first year. Third, spinoff activity can be sensitive to exogenous shocks that give rise to contagion-style market entry by a wide assortment of insiders, thereby generating the kind of performance heterogeneity (Hunt, 2013b, 2015) that is of research interest to strategy and entrepreneurship scholars. Fourth, there exists an increasingly well-structured explanatory framework for spinoffs, thereby providing the necessary structure to examine both the empirical and theoretical impacts of truncation (e.g. Klepper, 2009).

Illustrating Truncation Susceptibility: Entrepreneurial spinoff versus *de novo* performance

Central among the emerging "stylized facts" and "empirical regularities" regarding entrepreneurial spinoffs (Klepper, 2009) is the assertion that spinoffs live longer than *de novo* entrants (Agarwal, et al., 2004; Klepper, 2009). The logical basis for this is the view that "spin-outs have a survival edge in the market over other entrants as the result of a combination of entrepreneurial flexibility and inherited knowledge" (Agarwal *et al.*, 2004: 519). Core assertions of this nature are essential to the explanatory framework for spinoffs; but, because they hinge on early-stage developments in the life cycle of nascent firms, they are susceptible to truncation effects. For this reason, data comparing longevity for spinoff and *de novo* entrants is a constructive context for an examination of truncation effects (Hunt 2013b, 2015).

Paradoxically, a test of truncation involves an evaluation of that which is unobservable. Therefore, the most useful contexts through which to illustrate truncation consist of industry populations for which the presence of missing data is impossible, or highly unlikely. This requires circumstances in which the presence of an operating entity is mandated by law, such that it not possible to offer goods and services in a given sector without a formally defined presence. Fortunately, the asbestos abatement industry meets these stringent requirements. Beginning in 1986, U.S. law required the removal and disposal of asbestos-containing materials by annually licensed firms, employing annually trained and licensed abatement professionals, completing work through of project-specific, State-issued permits (Hunt 2015).

Equipped with this level of comprehensive detail, the test of left-side truncation effects can be carried out in two respects. First, using milestones that are commonly used in strategy research – such as, venture financing, appearance in catalogues, appearance in trade journals, or operational visibility – we ask: what are the statistical effects of truncation in the context of a complete population of industry participants? Second, we pose the question: what theoretical challenges arise from evidence of statistically significant truncation? In the following section, we explore both of these questions in the context of a complete industry.

The asbestos abatement context

Leading up to the 1980s, the use of asbestos was extensive. Unfortunately, when disturbed, it is harmful to humans. In response, Congress passed the Asbestos Hazards Emergency Response Act (AHERA) in 1986, establishing standards requiring professional abatement of asbestos in existing structures, giving birth to the asbestos abatement industry. Functionally, the enforcement of AHERA was delegated to the State-level. Colorado, a high-enforcement state, maintains a comprehensive archival record, available through the Colorado Department of Public Health & Environment (CDPH&E), covering the industry from its inception in 1986.

As a consequence of the monitoring and reporting requirements for the removal and disposal of asbestos containing material, extensive industry, firm and individual data are available. These data allow the capture of all firms, even those that fail to complete even one project or survive beyond their first annual license. Since individual testing and certification renewals are required annually and lists are publicly available, new firms and the individuals who found them can be continuously tracked through CDPH&E and Colorado Department of State records. For example, in 2001 there were 24 new industry entrants: 16 of these were spinoffs, firms founded by former employees of existing abatement firms; 6 of the new entrants were *de novo* firms, founded by individuals new to the industry; and, 2 were *de alio* entrants, consisting of existing firms diversifying into asbestos abatement from some other industry. Through 2010, 612 firms had entered the industry and 508 had exited. 448 of the entrants, representing just over 70%, were spinoffs (Hunt 2013b).

Dependent Variables, Predictors, and Model Specifications

In order to investigate the potential significance of truncation effects, we developed models that measured lifespan and operational performance as a function of entry mode (spinoff versus *de novo*) under both truncated and non-truncated conditions. The dependent variable, *Firm Lifespan,* refers to the total duration of operational existence measured in years. Our key predictor for this comparative analysis is *Entry Mode,* a dichotomous variable with 1 indicating an entrepreneurial spinoff, and 0 indicating *de novo* entrants. This variable was based on whether the firm founder came from an existing firm within the industry. Additionally, we employed macroeconomic, industry-specific and firm-specific control variables. The macroeconomic vector contains the State-level measures for construction, unemployment, and economic activity. Our model also controls for unobservable fixed-year effects. Industry-specific measures consist of selected predictors substantiated by organizational ecology (e.g. Hannan & Carroll, 1992): density at birth entry cohort size and average entry cohort lifespan. To account for changes in the industry scale, we included total projects completed. Firm-specific measures consisted of founder experience and the average number of projects completed.

To simply and clearly illustrate the threat and effect of LST through interpretable coefficients, firm survival was modeled using an OLS regression model represented by:

$$Firm\ Lifespan = \beta_0 + \beta_1 CON_{macro} + \beta_2 CON_{indus} + \beta_3 CON_{firm} + \beta_4 ENTRYMODE \tag{1}$$

Empirical findings

As noted from the outset, theoretical frameworks aiming to characterize entry and survival strategies are susceptible to pronounced truncation effects (Yang & Aldrich, 2012). Table 2 displays the comparison between the truncated and non-truncated populations as a function of entry mode.

[INSERT TABLE 1 ABOUT HERE]

Consistent with the relationships illustrated in Figure 1, it is apparent that truncation effects do in fact lead to an understatement of market entry and an overstatement of average performance, thereby affecting both of the metrics that are critical to a complete and thus accurate characterization of entry activity and subsequent operational outcomes. Spinoff entry examined under truncated conditions is understated by 40% relative to the actual, known, non-truncated population. Meanwhile, average longevity is overstated by 152% when the firms failing to survive longer than one year are truncated. These material differences stem from a condition commonplace in strategy studies: the "evaporation" of data associated with firms that never became substantively operational and the derivation of truncated means from the pool of remaining observable survivors.

Further substantiation of truncation effects is elaborated in Table 2, which shows the results of the comparison between the truncated and non-truncated populations. Model 1 represents the complete industry population of 558 spinoffs and *de novo* entrants, while Model 2 represents a truncated dataset of 345 firms reflecting that entrants failing to survive even a year would not have become substantively operational and would unlikely be unobserved. In other words, Model 2 reflects the commonplace reality that firms failing to survive even one year are often excluded from the pool of observable actors due to their limited operational presence.

[INSERT TABLE 2 ABOUT HERE]

As the results reveal, truncation not only distorts the regressed relationship between *Entry Mode* and *Lifespan*, but actually yields contradictory results in survival relative to *de novos*. While the truncated data (Model 2) results in a spinoff survival *advantage* of .36 years over *de novo* entrants, the non-truncated data (Model 1) results in a spinoff survival *disadvantage* of 1.24

years. The sign reversal between Models 1 and 2 for the coefficients of the predictor *Entry Mode* is tantamount to a theoretical fissure; it suggests that contrary to extant predictions, being a spinoff is not advantageous, but rather appears to be a significant liability when examined within the context of a complete industry population. Stated differently, when quickly failing, non-operational entrants are missing from a dataset (Model 2), an *apparent* spinoff advantage is seen. Yet, if able to observe all entrants, the contrary is found.

A Comparison of Existing Approaches for Left-Side Truncation

As the foregoing results indicate, a significant disconnect exists between the actors, events, and outcomes that are observed versus those that are typically unobserved. For this reason, the "missingness" of the unobserved is non-ignorable When missing data are non-ignorable a correction is necessary in order to avoid producing spurious results, misspecifications or faulty support for an underlying theory. We tested six existing approaches commonly applied to the missing data problem. The comparative results are presented in Table 3.

[INSERT TABLE 3 ABOUT HERE]

In each case, we ran the truncated regression model (Table 2, Model 2c) in SPSS, utilizing each approach in the fashion classically conceived in illustrative, representative applications of the techniques. As Table 3 reveals, each of the six approaches is problematic in one or more respects. Of the six, only the power law distribution approach (e.g. Crawford, et al., 2015) is partially accurate in noting the potential inadequacy of the truncated sample. The other five approaches find, quite incorrectly, that not only is the truncated dataset efficacious, but also the more egregious conclusion that the truncated Model (2c) and non-truncated Model (1c) are statistically indistinguishable. None of the five are able to catch the sign reversal for *Entry Mode* that occurs between the two models. This mistake includes the results of the Heckman two-step procedure, which involves generating an Inverse Mills Ratio that is subsequently included in the regression model. The two-step procedure is the most prevalent approach to correcting for survivor's biases among strategy and entrepreneurship scholars (Crawford et al. 2015). In conventional instances of data missingness, wherein missing data substantively resembles that which is in-hand, the Heckman two-step procedure is indeed a reliable tool. In this case, however, the missing data does not bear sufficient resemblance to the survivor data, rendering efforts to extrapolate from the existing data seriously and demonstrably inaccurate.

As for approaches employing power law distributions, the procedure correctly signals that the *Entry Mode* variable is susceptible to misspecification in the truncated model (2c); yet the power law approach fails to catch the statistically significant sign reversal in *Entry Mode* between the truncated and non-truncated models. We suspect, therefore, that power law-based corrections may be insufficiently attuned to the specific conditions that relate observed to non-observed events. We offer the means to alleviate that shortcoming in the next section.

REPORTING A TRUNCATION FACTOR

Absent the rare events giving rise to comprehensive datasets and natural experiments, how can strategy scholars address the perils of truncation? Applying a sensitivity analysis approach (Daniels & Hogan, 2008), we propose that a readily implementable *Truncation Factor* should accompany the use of archival datasets, indicating the potential missing data bias. Analytical reporting of robustness thresholds has precedence in organizational research. For example, single-source, self-report data that may be subject to common method bias employs *post hoc* statistical techniques to detect, assess and report on such biases (Podsakoff *et al.,* 2003). Even more closely related is the "file drawer" issue in meta-analytic studies. As a robustness check in meta-analyses, Hunter and Schmidt (2004) and Rosenthal (1984) advocate *post hoc* reporting involving the calculation of the number of null-effect studies that could be included before calling into question the significance of the meta-analytic findings. In analogous fashion, archival studies subject to possible data truncation (Shen, 2005) would be well-served by the use of a *Truncation Factor,* indicating the extent to which findings are stable in the presence of non-random, non-ignorable missing observations.

An informative *Truncation Factor* would report the maximum non-conforming data (i.e. unobserved data that are inconsistent with the truncated results) that can be added to the expositional model while keeping the statistical significance of the relationships intact. The *Truncation Factor* we propose is calculated using the logic displayed in Equation 2 below. (***L***) provides a first step in gauging the threat posed by left-side truncation as it represents the maximum percentage of non-conforming events that could be added to the observed events without losing the statistical significance of the results. (***K)*** is an estimate for the base-rate occurrence of non-conforming events pertinent to the line of inquiry and determines the probability that ***L*** will overwhelm the truncated results:

$$\textbf{Truncation Factor} = \frac{\mathbf{K}_{\textbf{Base Rate}}}{\mathbf{L}_{\textbf{Model Maximum}}} \tag{2}$$

To the greatest extent possible, ***K*** should represent an unbiased estimate of the prevalence of contra-indicated events relative to the hypothesized relationships. For example, in the case of start-up survival, Shane (2008) found that 25% of all the new firms tracked by the U.S. Small Business Administration typically fail in the first year. This metric offers an evidence-based, independent ***K*** for the expected level of early firm failures that might be subject to left-side truncation in an analysis of spinoff survival and performance. The value for ***L*** in our truncated population in Model 2 (Table 2) was calculated to be 9%. This was derived by adding early failures to the analytical pool until the focal coefficient for Entry Mode in Model 2 was no longer significant at $p < .05$. This equates to 31 non-conforming data points (i.e. spinoff entrants that failed before reaching the observation window) that could be added to the truncated model without losing significance. Absent an independent basis for an unbiased ***K***, the ***L*** can be reported as well as Truncation Factors based on a range of possible ***K***s derived through sensitivity analysis

The relationship between ***K*** and ***L*** is an empirically derived assessment of the threat posed by truncation. If ***L*** (the maximum allowable number of missed, non-ignorable observations relative to total observations) is greater than ***K*** (the estimated base-rate of non-conforming events in the true population), a Truncation Factor < 1.0 results. This suggests that the findings are relatively resistant to truncation effects since the estimated prevalence of non-ignorable events is less than the number of non-conforming events that are needed to nullify the statistical significance of the base case model. On the other hand, if the value of ***L*** is less than ***K***, it produces a TF > 1.0, implying that the empirical results are relatively sensitive to truncation. Put another way, a TF greater than 1.0 suggests that the missing data is apt to be **non**-ignorable, threatening material distortions of the research findings.

In relation to the previously reported spinoff findings/datasets, the following *TFs* can be observed. Applying Equation 2 to Model 2 in Table 2, a TF of 2.78 is obtained (25% / 9%, or 0.25/0.09), indicating that Model 2 is highly susceptible to truncation effects. The findings are at risk because the independent predictor of non-conforming events is much larger than the maximum number of non-conforming events that the base model can tolerate. Since a 25% first-year failure rate (Shane, 2008) is the independent estimate for ***K***, a maximum threshold of 9% for

L means that the model can only tolerate 36% of the non-conforming events that are likely to have been unobserved (i.e. 9% / 25%, or 0.09/0.25). Indeed, this is precisely what we found when the Model 2 sample is compared to the non-truncated Model 1. The apparent spinoff survival superiority of 0.36 years in the truncated Model 2 gives way to survival inferiority of 1.24 years in the non-truncated Model 1. Thus, the TF of 2.78 correctly warned that Model 2 is highly susceptible to the adverse effects of non-ignorable missing data; in this case, it signaled a high risk that Model 2 understates spinoff entry and overstates spinoff survival.

The same test can be applied to the complete population of firms of Model 1. That is, we can also calculate a TF for the comprehensive dataset (Table 2, Model 1), comprised of 558 firms. ***L***, in this instance is calculated by adding non-conforming, unobserved *de novo* failures to the model until the coefficient is no longer significant at $p < .05$. In this case, 317 such non-conforming events would have to be added. Thus, the TF = 0.44 (i.e. 25% / 57% or 0.25/0.57), with ***K*** = 25% (Shane, 2008) and ***L*** = 57% (i.e. 317/558). With a TF of 0.44, there is little chance that a sufficient quantity of non-conforming data was missed such that the findings derived from Model 1 are rendered insignificant. More concretely, the TF of 0.44 means that the study findings will not materially change unless 317 additional unobserved spinoffs entered the highly regulated abatement industry and failed before ever becoming licensed. With such an improbably high threshold -- since unlicensed abatement is illegal and readily observable -- it can be concluded that Model 1 is resistant to the confounding effects of left-side truncation.

CONCLUSION

As the foregoing analysis demonstrates, the effects of LST are empirically and theoretically impactful, echoing the view of Baum & Haveman (1997:304) that while "research has treated foundings as identical additions to homogeneous populations," in reality, unobserved organizations are likely to bear little resemblance to the observable pool when there exists a nonrandom, non-linear relationship between the successes and the failures. As a consequence, noted Yang and Aldrich (2012), there is a pervasive equivocality running through research results for new ventures. "Previous studies have been inconclusive regarding not only the patterns but also the magnitude of failure rates" (Yang & Aldrich 2012:478). This inconclusiveness arises when neither sampling remedies nor a wide array of statistical remedies succeed in generating a meaningful expression of the unobserved, non-linear relationship between long-term survivors and early failures. This line of analysis provides added weight to recent assertions that empirical findings related to early stage conditions, events and outcomes in strategy and entrepreneurship research may be hamstrung by normal distribution assumptions (Crawford et al. 2015). The various shortcomings in applying extant tools to the problem of LST (Table 3) make it difficult to compare empirical results across varied contexts. This, in turn, impedes attempts to build more robust explanatory frameworks for nascent-stage events by employing meta-analytic studies (Hunter & Schmidt 2004), since an unknown degree of truncation makes the synthesis of diverse datasets untenable.

The Truncation Factor that we have proposed offers a straightforward resolution to this impasse by creating a common basis for the identification and quantification of LST. Our illustration of entry mode and organizational survival suggests that Truncation Factor scores for academic studies of <1 could be viewed as having low susceptibility to LST since even a very large pool of conflicting data would not alter the findings. TF scores ranging from 1 to 2 indicate moderate levels of susceptibility. Arithmetically, explanatory models exhibiting this TF range could tolerate the effects of missing data representing a significant proportion of the observables.

TF scores of 2 or more represent more serious susceptibility to LST biases since even relatively small differences in the quantity and character of unobservable data could alter the results. In these instances, a study's findings may be subject to limitations and boundary conditions that should be prominently noted in published works. Regardless of the TF score for any given study, the practice of reporting TF services both the empirical and theoretical foundations of strategy and entrepreneurship research.

Since retrospective studies, even including those using incomplete archival data, have considerable value in addressing a wide array of important research questions, the remedy suggested by our study is not that truncation is so debilitating that studies using anything less than a complete population are untenable. Rather, the results illustrate: (a) unobservable events should not be assumed to resemble the observable events; (b) archive-based models should be counter-factually stress-tested using sensitivity analysis; and, (c) theories hinging on truncated data are potentially susceptible to non-ignorable missingness. Thoughtful implementation of a Truncation Factor can show robustness as well as provide an early-warning system that the adverse effects of truncation may require approaching a line of inquiry through alternative data or through an articulation of clear boundary conditions.

REFERENCES

Agarwal, R., Echambadi, R., Franco, A., & Sarkar M. 2004. Knowledge transfer through inheritance: Spinout generation, development, and survival. *Academy of Management Journal*, *47,* 501-522.

Amburgey, T., & Rao, H. 1996. Organizational ecology: Past, present, and future directions. A*cademy of Management Journal*, *39:5*, 265 – 86.

Andersen, P., Borgan. Ø., Gill, R., & Keiding, N. 1993. *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. Springer, New York.

Baum, J. A., & Haveman, H. A. (1997). Love thy neighbor? Differentiation and agglomeration in the Manhattan hotel industry, 1898-1990. *Administrative Science Quarterly*, 304-338.

Bilker, W., & Wang, M. 1997. Bootstrapping Left-Truncated and Right-Censored Data. *Communications in Statistics, Simulation and Computation, 26*, 141–171.

Cameron, A., & Trivedi, P. 2005. *Microeconometrics*. Cambridge Press

Chatterji, A. 2009. Spawned with a silver spoon. *Strategic Management Journal*, *30,*185-206.

Crawford G, Aguinis H, Lichtenstein B, Davidsson P, & McKelvey B. (2015). Power law distributions in entrepreneurship: Implications for theory and research. *Journal of Business Venturing*, 30(5), 696-713.

Daniels, M., & Hogan, J. 2008. *Missing Data in Longitudinal Studies*. Boston: CRC Press.

Delacroix, J., & Carroll, G. 1983. Organizational foundings: An ecological study of the newspaper industries of Argentina and Ireland. *Administrative Sci Quarterly*, 28*,* 274 - 291.

Hannan, M. & Carrol, G. 1992. *Dynamics of Organizational Populations: Density, Legitimation and Competition*. New York: Oxford University Press

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica*, *47,* 153-161.

Helfat, C., & Lieberman, M. 2002. The birth of capabilities: market entry and the importance of pre-history. *Industrial and Corporate Change*, *11(4),* 725–760.

Hitt, M., Gimeno, J., & Hoskisson, R. 1998. Current and future research methods in strategic management. *Organizational Research Methods*, *1,* 6-44.

Hunt, R. A. (2013a). Entrepreneurial tweaking: An empirical study of technology diffusion through secondary inventions and design modifications by start-ups. *European Journal of Innovation Management*, *16*(2), 148-170.

Hunt, R. A. (2013b). Essays Concerning the Entry and Survival Strategies of Entrepreneurial Firms: A Transaction Perspective.

Hunt, R. A. (2013c). Priming the pump: demand-side drivers of entrepreneurial activity. *Frontiers of Entrepreneurship Research*, *33*(14), 2.

Hunt, R. 2015. Contagion Entrepreneurship: Institutional Support, Strategic Incoherence, and the Social Costs of Over-Entry. *Journal of Small Business Management*, 53(S1), 5-29.

Hunt, R.A., & Ortiz-Hunt, L. (2017). Entrepreneurial round-tripping: The benefits of newness and smallness in multi-directional value creation. *Management Decision*, *55*(3), 491-511.

Hunter, J., & Schmidt, F. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage Publications Inc, Thousand Oaks, CA.

Ketchen, D., Boyd, B., & Bergh, D. 2008. Research methodology in strategic management: Past accomplishments and future challenges. *Organizational Research Methods*, *11*, 643-658.

Kim, C., & Lai, T. 2000. Efficient score estimation and adaptive M-estimators in censored and truncated regression models, *Statistica Sinica 10,* 731–749.

Klepper, S. 2001. Employee startups in high-tech industries. *Industrial and Corporate Change*, *10,* 639-674.

Klepper, S. 2009. Spinoffs: Review and synthesis. *European Management Review*, *6,*159-171.

Little, R., & Rubin, D. 1987. *Statistical analysis with missing data*. New York: Wiley.

Podsakoff, P., MacKenzie, S., Lee, J., & Podsakoff, N. 2003. Common method biases in behavioral research. *Journal of Applied Psychology*, *88*, 879-903.

Roderick, J., & Rubin, D. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Rosenthal, R. 1984. *Meta-Analytic Procedures for Social Research*. Sage, Beverly Hills, CA.

Shane, S. 2008. *The Illusions of Entrepreneurship: The Costly Myths that Entrepreneurs, Investors, and Policy Makers Live By.* Yale University Press: New Haven, CT.

Shen, P. 2005. Estimation of the truncation probability with left-truncated and right-censored data. *Journal of Nonparametric Statistics*, *17(8),* 957 – 969.

Short, J., Ketchen, D., Combs, J., & Ireland, R. 2010. Research Methods in Entrepreneurship Opportunities and Challenges. *Organizational Research Methods*, *13,* 6-15.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica*, *26* (1): 24–36.

Tsiatis, A. 2006. *Semiparametric Theory and Missing Data*. Springer Series on Statistics.

Yang, T. & Aldrich, H. 2012. Out of sight but not out of mind: Why failure to account for left truncation biases research on failure rates. *Journal of Business Venturing*, *27*(4), 477-492.

Figure 1: Stylized Schematic of the Data Truncation Effect.

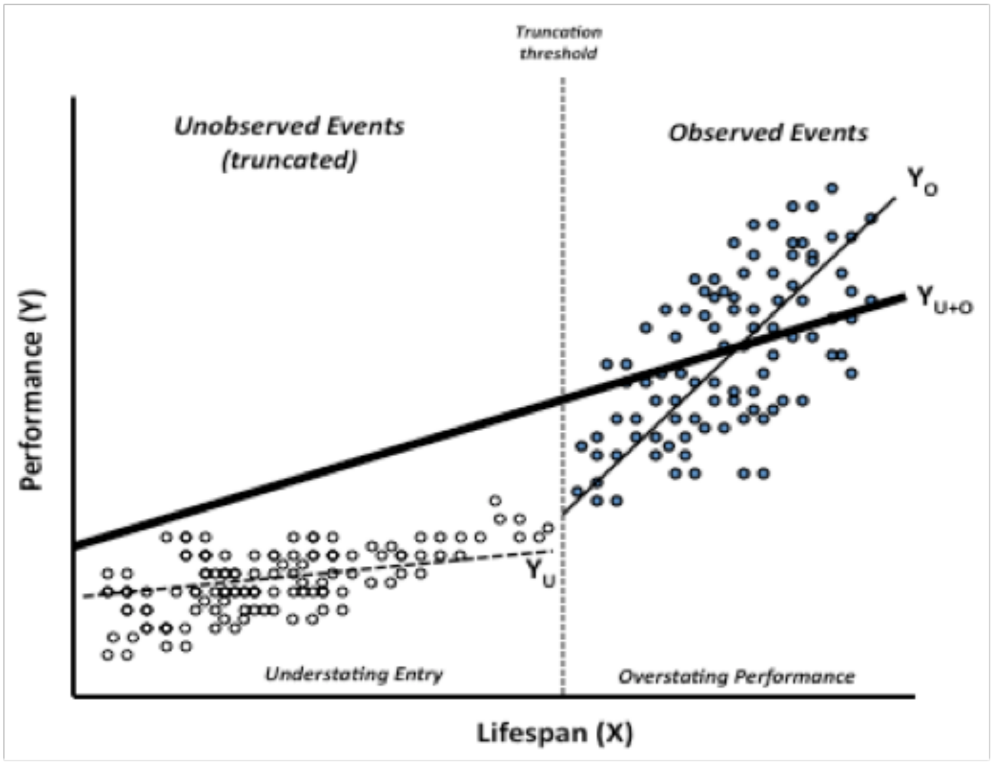

Table 1: Truncation Comparison -- Spinoff and De Novo Entry

	Non-Truncated (N = 558)			Truncated (n = 345)‡		
Entry Mode	**# Firms**	**Avg. Lifespan (yrs)**	**Avg. Op. Performance (projects/yr)**	**# Firms**	**Avg. Lifespan (yrs)**	**Avg. Op. Performance (projects/yr)**
Spin-off	448	3.1	18.1	270	7.8	41.2
De Novo	110	5.6	27.4	75	6.8	37.3
Total Firms	558	3.7	23.8	345	7.4	40.4

‡ Firms surviving at least one year.
All focal mean differences highly significant, p < .001.

Table 2: Results of OLS Estimation for Firm Lifespan

Predictor	Models					
	Non-Truncated Data (n = 558)			**Truncated Data (n = 345) ‡**		
	1a	**1b**	**1c**	**2a**	**2b**	**2c**
(Constant)	4.635***	4.737***	5.856***	4.113**	4.464***	4.935**
	(1.066)	(.906)	(1.113)	(.824)	(.807)	(.741)
Macro Controls	-0.034	-0.031	-0.022	-0.102*	-0.093	-0.091
	(.013)	(.010)	(.007)	(.117)	(.110)	(.110)
Density at Birth	-0.021**	-0.022*	-0.023*	0.142*	0.137*	0.131*
	(.011)	(100)	(.011)	(.457)	(.452)	(.448)
Entry Cohort Size	0.002	0.002	0.018	(-.089)*	(-.033)*	(-.024)*
	(.028)	(.028)	(.028)	(.031)	(.017)	(.024)
Cohort Lifespan	0.011	-0.009	0.001	-0.001	-0.001	0.000
	(.025)	(.030)	(.025)	(.002)	(.001)	(.001)
Founder Experience		0.077*	0.049*		0.136	0.122
		(.051)	(.027)		(.034)	(.029)
Year of Entry		-0.079	-0.074		0.084	0.079
		(.060)	(.057)		(.055)	(.052)
Avg Annual Projects		0.144***	0.130**		0.175**	0.169**
		(.021)	(.013)		(.018)	(.015)
Total Projects		0.114***	.112***		0.160**	0.11**
		(.001)	(.001)		(.002)	(.001)
Entry Mode (1 = Spinoff)			**-1.242*** **			**0.362***
			(.361)			**(.194)**
Adj. R2	0.331	0.585	0.772	0.414	0.534	0.591
F-value	34.7***	111.2***	118.8***	50.8***	80.9***	57.7***

‡ Models only those entrants existing for at least a year.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3: A Comparison of Existing Approaches to Left-Side Truncation

Approach	Illustrative Uses	Application of Remedy to Asbestos Industry LST	Outcome
Two-Step Procedure	Heckman (1979)	Inverse Mills Ratio provides support for truncated regression model.	Incorrectly finds truncated model (2c) to be accurate and indistinguishable from the non-truncated model (1c).
Estimating hazard of early termination	Yang & Aldrich (2012)	Assumes missing data substantively resembles available data.	Incorrectly finds truncated model (2c) to be accurate and indistinguishable from the non-truncated model (1c).
Power Law Distribution	Crawford, Aguinis, Lichtenstein, Davidsson & McKelvey (2015)	Successfully accounts for nonignorability of missing data, but over-corrects even when truncated data is made available.	Detects potential problems with truncated model, but fails to differentiate between truncated and non-truncated datasets.
Partial Likelihood Approach	Andersen, Borgen, Gil & Keiding, (1993)	Assumes availability of entire covariate history so that available data is assumed to resemble missing data.	Incorrectly finds truncated model (2c) to be accurate and indistinguishable from the non-truncated model (1c).
Boot-Strapping	Bilker & Wang (1997)	Resampling based on the assumption that existing data is normally distributed and resembles unobservables.	Incorrectly finds truncated model (2c) to be accurate and indistinguishable from the non-truncated model (1c).
Semi-Parametric	Shen (2010); Tsiastis (2006)	Relaxes parametric requirements to accept range of potential solutions. This results in misspecifying the truncated predictors, particularly *Entry Mode*.	Incorrectly finds truncated model (2c) to be accurate and indistinguishable from the non-truncated model (1c).