EMBRY-RIDDLE Aeronautical University SCHOLARLY COMMONS

International Journal of Aviation, Aeronautics, and Aerospace

Volume 4 | Issue 4 Article 9

11-12-2017

Using Data Envelopment Analysis to Benchmark Safety Culture in Aviation Organizations

Benjamin J. Goodheart Ph.D.
Versant / Embry-Riddle Aeronautical University Worldwide, bgoodheart@versantrisk.com

Using Data Envelopment Analysis to Benchmark Safety Culture in Aviation Organizations

Cover Page Footnote

The author acknowledges with gratitude the efforts of Robin Reidel and Samson Fatokun in reviewing this research and providing their feedback to ensure its accuracy, readability, and practicality. Thank you both for your help.

With safety management systems (SMS) occupying an increasingly prominent position in the aviation landscape, the ability to benchmark safety performance not only against past internal effectiveness but also across similar organizations seems a necessary element of one of SMS's foundational themes: continuous improvement. At present, anecdotal evidence, supported by limited available research in the field, suggests that few organizations are engaged in meaningful evaluation and benchmarking of safety performance through an established methodology. This research investigates a method of evaluating an organization's efficiency in creating a robust and positive safety culture using efficiency frontier estimation, specifically through data envelopment analysis (DEA).

Introduction

Whether as a byproduct of legislative requirement or as a result of corporate obligations and responsibility, aviation operators are increasingly implementing structured safety systems, many in the form of SMS. One of the "essential" elements of an SMS is a "safety culture" (Federal Aviation Administration, 2015, pp. 3-4). Acknowledging that the notion of safety culture has become widely accepted, if not comprehensively defined, the International Civil Aviation Organization (2013) addresses culture in the context of safety as largely an organizational issue that is supported by and manifested in safety reporting procedures and practices designed to identify inherent hazards. Safety management systems seek to create an environment in which organizations move from mere compliance with minimum regulatory requirements to a performance-based approach (International Civil Aviation Organization, 2009). In a performance-based safety environment, there must be measurable indicators of the capacity for safety beyond lagging indicators such as accident or injury rate. The logical inference, consistent with the axiom often attributed to the management consultant Peter Drucker: "if you can't measure it, you can't manage it," is that safety culture should be the subject of consistent evaluation and measurement as part of attempts to purposefully manage or shape an organization's culture. This assertion is supported by standards such as those promulgated in the International Standard for Business Aircraft Operations (IS-BAO), ISO 9001:2000 and ISO 9001:2008, and AS 9100, all of which contain language referencing the achievement of objectives and targets, performance measurement and monitoring, and continual improvement; elements which are arguably central to the concept of safety culture.

Despite the apparent requirement for a quantitative means of measuring and benchmarking safety culture and performance in aviation organizations, a commonly accepted approach does not presently exist. This research addresses this

problem by proposing the development of a mathematical model through the use of data envelopment analysis (DEA) to provide a quantitative benchmark of efficiency in establishing safety culture relative to a set of organizations, or from a single organization based on historical data.

This research contributes to the body of knowledge by introducing a methodology for assessment of the relative efficiency of organizations in creating and maintaining a positive safety culture, and as such, the relative effectiveness of safety programs. The results of this study, when applied to operational data, present a method to guide organizations in identifying the most appropriate means of increasing their efficiency in developing a thriving safety culture. It further forms a foundation on which additional investigation of stochastic measurements, such as bootstrapped DEA, can be adopted as a measurement system.

Review of Literature

Safety Culture Measurement

Measurement of safety program effectiveness is an onerous task for many organizations, not only those in the aviation industry. In the context of safety, traditional performance indicators tend toward a focus on forensic data, often reporting metrics like accident and incident rate, cost of accidents, insurance claims, and lost days due to injury (Arezes & Miguel, 2003). Peterson (2001) argues that safety systems should be evaluated based on a multiple-measures approach, suggesting that safety programs should be assessed using a minimum of three measures: accident record; audit scores, and results from perception surveys. In this context, perception surveys are intended to measure the state of an organization's safety culture, an issue which has been the subject of growing attention within the aviation community (Wiegmann, Zhang, von Thaden, Sharma, & Mitchell Gibbons, 2004).

Understanding safety culture as a science is somewhat of a paradox, given that it can act "simultaneously as precondition both for safe operations and for the oversight of incubating hazards" (Pidgeon, 1998, p. 205). This observation is suggestive of the somewhat abstract status of safety culture as a concept in practice. It is generally accepted that safety culture operates as a subset of a larger organizational culture in the same context as identified by Zohar (1980), though consensus has not necessarily been reached on exactly what the concept is, or on why it should be the subject of empirical inquiry (Frost, Moore, Louis, Lundberg, & Martin, 1991). Guldenmund (2010, p.1466) notes that "culture is an intangible, fuzzy concept encompassing acquired assumptions that is shared among the

members of a group and that provides meaning to their perceptions and actions and those of others." Schein (2010) generally agrees with Guldenmund's characterization of the concept, despite also suggesting that there is no such thing as a safety culture independent of the more significant organizational construct (Conklin, 2016). Despite an apparent lack of conceptual specificity, from a practical standpoint, the study of safety and its related constructs is of apparent importance.

Beus, Payne, Bergman, and Arthur (2010) cite 2008 U.S. Bureau of Labor Statistics figures that indicate in the United States in 2007; there were over 5,600 work-related fatalities and over four million nonfatal work injuries. In the environment of aviation operations and specifically those that occur during ground operations, the Flight Safety Foundation (n.d.) estimates that around 27,000 ramp accidents occur worldwide each year, with over 243,000 injuries and a total cost of at least $10 billion annually. Although they are illustrative as examples of why safety endures as an area of intense scrutiny, these figures make no provision for attempting to capture the psychological costs or damage to an organization's reputation as the result of an accident (Neal & Griffin, 2002).

Attempts at measuring safety culture have come about in part as a result of the realization that the forensic, retrospective measurements that have long characterized attempts to assess safety, such as accident or incident rate, injuries, or fatalities, provide little support for proactive initiatives (O'Connor, O'Dea, Kennedy, & Buttrey, 2011). In aviation, as in many industries, actual accident rates have fallen to so low a level as to make their analysis as a measure of safety performance almost obsolete. In fact, the measurement of such metrics as accident rate can be viewed mainly as a reactive response to events that, assuming a functional safety system, theoretically occur with decreasing frequency (Keren, Mills, Freeman, & Shelley, 2009). It is often difficult to discern, reliant on such low base-rate phenomena, whether a lack of injuries or accidents is the result of organizational measures designed to reduce hazards, or if the absence of injury is merely a reflection of too short a period of observation (Beus, Payne, Bergman, & Arthur, 2010). Rather than require that a system reach failure as a means of identifying latent faults, the concept of measuring safety culture as a leading indicator of organizational safety performance is appealing from a management perspective in that it allows the potential for reduction of operational accidents through a proactive, and even predictive, approach.

While it is generally understood at the theoretical level that a positive safety culture is an essential element in the prevention of accidents, Wiegmann et al. (2004) note that no broadly standardized tools exist that can be used to assess safety culture. Reviewing the extant literature on the subject, however, Wiegmann

et al. (2004) identify five common themes that serve as indicators of safety culture. These indicators include organizational commitment, management involvement, employee empowerment, reward systems, and reporting systems. The importance of organizational commitment is echoed by Flin, Mearns, O'Connor, and Bryden (2000), who identified management behaviors and attitudes as an emergent theme in their own research. Fuller (1997) also addresses organizational commitment as being a key indicator of safety performance, though not in an aviation-specific occupational health and safety context. In reviewing assessment instruments in the UK offshore oil industry, Davies, Spencer, and Dooley (2001) found that of 114 different questionnaire items, 30 related directly to an organizational and managerial commitment to safety. Management involvement, like organizational commitment, is also well-supported as an indicator of the health of an organization's safety culture. Choudry, Fang, and Mohamed (2007) assert that management commitment to safety initiatives and compliance is central to the very definition of safety culture, an observation echoed by Flin et al. (2000). That comparison of a different culture or climate assessment tools would indicate a preponderance of questions centered about managers' safety behaviors is not surprising, given that supervisors "undoubtedly set the tone and tempo for organizational atmosphere" including perceptions regarding safety (Flin et al., 2000, p. 186).

Employee involvement appears in a number of safety culture instruments (Davies, Spencer, & Dooley, 2001; Fuller, 1997) and its importance is stressed by Geller (1994), who notes that employees who feel "part of a cohesive group" are more likely to act empathetically toward their fellow employees, increasing safety-compliant behavior (p.21). Reason (1997) speaks to the matter of rewards, commenting that rewarding desirable safety behaviors can be a key element of creating a just culture, which he argues is a necessary component of a functional safety culture. Reward programs, naturally, rely on the consistent use and formal documentation; but when these conditions are satisfied, properly designed and implemented rewards can positively reinforce safety behaviors and safety culture (Wiegmann et al., 2004). Finally, the assertion by Wiegmann et al. (2004) that a healthy reporting system is a key indicator of a safety culture is echoed by Reason (1997), who notes that the rationale for any reporting system is to identify "organizational factors promoting errors and incidents" (p. 198). In summary, clear evidence exists within the literature in support of the previously identified indicators of safety culture performance.

The body of knowledge of safety culture clearly suggests that the construct of safety culture is still developing, with a litany of definitions and tools for assessment. However vaguely-specified the concept of culture may be, the evidence

overwhelmingly supports its importance as a foundation for positive safety performance. The literature is also far from ambiguous in its support of attempts to quantify and analyze measures of safety culture as a leading indicator, allowing organizations to assess safety performance even in the absence of accidents. Data envelopment analysis is one such method for evaluation and analysis of safety culture

Use of DEA for Benchmarking

DEA is a robust, nonparametric alternative to regression analysis for measuring the efficiency of business operations as compared to an estimated production efficiency frontier (El-Mashaleh, Rababeh, & Hyari, 2010; Ray, 2004). This frontier is the estimated, but unobserved "geometrical locus of optimal production plans" (Simar & Wilson, 1998, p. 49). The estimated technical efficiency of any decision-making unit is the ratio of the distance from the origin to the unit under evaluation and the distance from the origin to the composite unit on the efficiency frontier (Barros & Dieke, 2008). Introduced by Farrell in 1957 and refined to form the techniques which serve as the basis for this research by Charnes, Cooper, and Rhodes in a 1978 article, DEA was proposed as a means of objectively evaluating the efficiency of each participating decision-making unit (DMU) in a public program. In its original form (hereafter referred to as the CCR model after Charnes, Cooper, and Rhodes), DEA was limited by an assumption of constant returns to scale; that is, outputs change proportionately to a change in all inputs (Charnes et al., 1978). Following up on their earlier model, Banker, Charnes, and Cooper (1984) introduced a more refined set of variables to the DEA ratio that allowed for variable returns to scale, accommodating contestable market theory rather than a simple, single output economic model.

The ability of the Banker, Charnes, and Cooper (BCC) model to accommodate reallocation of resources to more attractive or more profitable activities allows for the possibility that resources available within many safety systems are dynamic and may be apportioned as such. For this research, however; the inability of the BCC model to accommodate inefficiencies beyond those attributed to technical or managerial elements makes it an inappropriate technique as compared to the CCR model, which includes estimation of scale efficiencies, combining estimates into a single value while assuming constant returns to scale (Barros & Dieke, 2008). As opposed to methods such as stochastic frontier analysis (SFA), DEA was chosen in this study because, unlike in SFA, it requires neither a large sample size nor that the functional form for estimation of cost or production technologies be pre-specified, a characteristic that limits flexibility and makes SFA ill-suited to the research at hand (Assaf & Josiassen, 2011).

El-Mashaleh, Rababeh, and Hyari (2010) undertook a study to benchmark the safety performance of construction contractors in Jordan. The El-Mashaleh et al. (2010) benchmarking study sought to compare the efficiency of utilization of safety-related expenses with safety expenses as a percent of total revenue as input variables and accident count as output variables. The authors acknowledge a limitation of their technique implicitly by suggesting a separate analysis of variance (ANOVA) be utilized to further examine differences between contractors. This study demonstrates that safety performance can be benchmarked against homogenous operators as well as internal measures evaluated over time through the use of DEA. In a similar vein, Beriha, Patnaik, and Mahapatra (2011) studied the applicability of DEA as a tool for the evaluation of safety performance in Indian organizations in the construction and manufacturing sectors.

The use of DEA in production and technical efficiency estimation is due in large part to its ability to compare DMUs directly to peer organizations while avoiding the many assumptions inherent in similar techniques such as stochastic frontier analysis (Ray, 2004). Adding to the applicability of DEA as a tool for safety benchmarking is that inputs need not be expressed as homogenous units; a trait that is particularly appropriate given that influences of safety performance may occur across varying levels of measurement. Despite its apparent applicability to benchmarking, DEA is not without its limitations. First, because it is an extreme point technique, DEA is particularly susceptible to measurement and sampling errors (Simar & Wilson, 1998). Second, DEA as a method is only concerned with performance relative to the sample. If for instance, none of the subject organizations is efficient relative to a theoretical maximum value, DEA will only reflect relative efficiency and not the gap between reality and how well an organization could ideally be performing. Finally, and of particular interest in this study, DEA is considered a nonparametric, or deterministic, technique. As such, it does not produce standard errors and makes hypothesis testing extremely difficult (Ray, 2004; Simar & Wilson, 1998). With this in mind, the present study was designed as a first step toward demonstrating the applicability of DEA to safety culture benchmarking in aviation, with the secondary goal of illustrating the need, along with suggestions for future directions, for expansions to native DEA techniques that allow stochastic efficiency measures.

Method

This research assumed a two-phase approach for model development. In the first phase, a native, nonparametric DEA model was developed and tested. In the second phase, the primary, native DEA model was used as a basis for the limited expansion of the model into stochastic, double-bootstrapped DEA form. In this

study, each organization whose measurement data are used as input information for DEA was considered a DMU. In a parallel use of the model, it is possible for an individual organization to self-audit and track performance at regular intervals considering each measurement taken over time as a separate DMU. In order to avoid inconsistencies in the measurement of production output, in this case, the previously-discussed indicators of safety culture, this study proposes to use a single conventional cultural measurement instrument for all DMUs. For the purposes of this research, the cultural survey described and implemented by von Thaden, Wiegmann, Mitchell, Sharma, and Zhang (2003) has been identified for its harmony with the five aforementioned cultural traits. To ensure that input and output variables are correctly identified as well as relevant to the intended efficiency estimate, the methodology outlined by Simar and Wilson (2001) was employed.

Variables in a DEA model represent the conversion of inputs, such as resources or investment, into outputs, such as specific performance measures by the DMUs under study (El-Mashaleh, Rababeh, & Hyari, 2010). In the context of the present research, input and output variables are different from those that typify DEA models which aim to estimate financial or production efficiency. In early stages of this research, a number of variables for potential inclusion in the model were identified, including: annual safety budget (per capita or as a percentage of total budget); normalized lost-time injuries; normalized aircraft incidents; normalized hazard report submissions; mean-time-to-resolution of hazard reports; count of overdue hazard resolutions; workers' compensation experience modifier; aircraft and general liability insurance loss ratio; and standardized scores derived from the selected cultural measurement instrument or its subsections. While the inclusion of many input or output variables could arguably increase the fidelity of the measurement, one limitation of DEA in this regard is the tendency for the number of variables to be too great relative to the sample size (Ray, 2005).

To alleviate this concern, variables were checked in accordance with the methods proposed by Jenkins and Anderson (2003) for confirmation of input or output status and were aggregated to the extent practicable. The simplified output variable is the composite score of the von Thaden et al. (2003) cultural measurement instrument previously discussed. Input variables were limited to include measures selected to represent three facets of investment: capital, personnel, and leadership. Each of these variables relates back to the factors addressed by von Thaden et al. (2003). Table 1 summarizes the variable definitions and assumptions used in the context of this research.

Table 1
Variable Descriptions and Measures

Variable	Input/ Output	Description	Measure
Percent of total annual expenses spent on safety	Input	Measures the financial commitment of the organization to safety. The higher the metric, the higher the share of its expenses that an organization devotes to safety; and thus, the higher the financial commitment to safety. Computed as the ratio of annual expense on safety to the total expenses of the organization.	Generally, the metric will be computed from actual annual expenditures. However, it may also be computed from budgeted annual expenses.
% FTEs dedicated to safety positions	Input	This metric measures the human resource commitment of the organization to safety. The higher the metric, the higher the share of total workforce that an organization devotes to safety and thus, the higher the human resource commitment to safety. Computed as the ratio of the total full-time equivalents (FTE) dedicated to safety to the total FTEs of the organization.	The use of FTE accounts for resources partially dedicated to safety, such as a pilot who in addition to flying duties also serves for 10% of work time on the union's safety committee. Such a resource would be counted as .1 of an FTE dedicated to safety. The metric can be computed for any given point in time (snapshot), or averaged over a period of time.
Mean hours of safety training per employee/year	Input	Measures the training commitment of the organization to safety. The higher the metric, the more safety training employees receive on average.	Computed as total hours of safety training received by employees in the organization during a year divided by the average total number of employees in the organization during that year.

Variable	Input/ Output	Description	Measure
Safety prominence in leadership communication	Input	This metric measures the organization's top leadership commitment to safety. The higher the metric, the more prominent safety is in top leadership communication.	Computed by scoring each top leadership message to employees for the prominence of safety, and averaging the scores over leadership messages sent that year. Scoring is based on the following point system: 0 for no mention of safety; 1 for single-line mention of safety (e.g., "Work safe and take care of each other), 2 for dedicating one paragraph to safety, 3 for dedicating more than one paragraph but not the entire message to safety, and 4 for dedicating the whole message to safety.
Safety culture	Output	von Thaden et al. (2003) culture assessment instrument.	Composite safety culture survey score.

The methodology proposed in this study has a strong foundation in the CCR model previously discussed, and it follows the linear programming model also used by Beriha, Patnaik, and Mahapatra (2011). This model, adapted from Charnes, Cooper, and Rhodes (1978), is a maximization of the ratio of weighted outputs to weighted inputs subject to the constraint that the equivalent ratios for every decision-making unit (DMU) are equal to or less than unity (p. 430). The input and outputs variables are represented by y_{rj} and x_{ij} and the variable weights determined by the solution are represented as u_r and v_i. This model is the basic foundation of DEA, and it aggregates measures of efficiency into a single value. There are two ways to state the ratio in the "ratio-form" or DEA model of technical efficiency (Cooper, Seiford, & Zhu, 2011). In the output-oriented form, the ratio considered is the ratio of outputs to inputs, where a higher ratio corresponds to higher efficiency. In the input-oriented form, the ratio considered is the reciprocal ratio, namely the ratio of inputs to outputs. In this case, a lower ratio corresponds to higher efficiency. For this research, the output-oriented CCR model was chosen as

a means of efficiency comparisons along a frontier based upon best practices or optimal implementation of cultural interventions.

The Beriha et al. model (2011) is a functional one and is nearly identical to the original CCR formulation. In the context of this study, the model is formulated in virtually the same arrangement, though written in ratio form as in Equation 1.

$$\max h_e(u,v) = \sum_{r\in R} u_r y_{re} \Big/ \sum_{i\in I} v_i x_{ie}$$

s.t.:

$$\sum_{r\in R} u_r y_{rd} \Big/ \sum_{i\in I} v_i x_{id} \le 1 \;\forall d \in D \tag{1}$$

$$u_r \ge 0 \;\forall r \in R$$

$$v_i \ge 0 \;\forall i \in I$$

Where: h_e is the objective function and represents the efficiency of DMU *e* which is evaluated,
u_r is a decision variable and represents the weight of output *r*,
v_i is a decision variable and represents the weight of input *i*,
y_{rd} is the observed value of output *r* for DMU *d*,
x_{id} is the observed value of input *i* for DMU *d*,
R is the set of all outputs,
I is the set of all inputs, and
D is the set of all DMUs.

To select a unique solution, the CCR model is transformed into a linear programming problem using the transformation by Charnes and Cooper (1962), which selects the representative solution in which the weighted sum of inputs equals unity. Equation 2 shows the resulting linear programming problem for the output-oriented CCR model.

$$\max E_e(\mu) = \sum_{r\in R} \mu_r y_{re}$$

s.t.:

$$\sum_{r\in R} \mu_r y_{rd} - \sum_{i\in I} v_i x_{id} \le 0 \;\forall d \in D \tag{2}$$

$$\sum_{i\in I} v_i x_{ie} = 1$$

$$\mu_r \geq 0 \; \forall r \in R$$
$$v_i \geq 0 \; \forall i \in I$$

Where: E_e is the objective function and represents the efficiency of DMU e which is evaluated,
μ_r is a decision variable and represents the weight of output r,
v_i is a decision variable and represents the weight of input i,
y_{rd} is the observed value of output r for DMU d,
x_{id} is the observed value of input i for DMU d,
R is the set of all outputs,
I is the set of all inputs, and
D is the set of all DMUs.

This linear programming model maximizes the sum of weighted outputs of the DMU that is being evaluated. The first constraint ensures that for each DMU the sum of weighted output is smaller or equal than the sum of weighted inputs. Through this constraint, a relationship between inputs and outputs is enforced. The second constraint ensures that the sum of weighted input for the DMU that is being evaluated is equal to one. Through this constraint, the uniqueness of a solution with maximum technical efficiency scores of one is enforced. The final two constraints ensure that the weights are non-negative.

The linear programming model shown in Equation 2 needs to be solved once for each DMU. To enable simpler processing of the model for the complete set of DMUs in commercial solver software, the researcher extended the model into a form in which the efficiency of all DMUs can be solved within one model run. Equation 3 shows the extended model.

$$\max \sum_{e \in D} E_e(\mu) = \sum_{e \in D} \sum_{r \in R} \mu_{re} y_{re}$$

s.t.:

(3)

$$\sum_{r \in R} \mu_{re} y_{rd} - \sum_{i \in I} v_{ie} x_{id} \leq 0 \quad \forall d \in D, e \in D$$
$$\sum_{i \in I} v_{ie} x_{ie} = 1 \quad \forall e \in D$$
$$\mu_{re} \geq 0 \quad \forall r \in R, e \in D$$
$$v_{ie} \geq 0 \quad \forall i \in I, e \in D$$

Where: E_e is the objective function and represents the efficiency of DMU e which is evaluated,
μ_{re} is a decision variable and represents the weight of output r for the evaluation of DMU e,
v_{ie} is a decision variable and represents the weight of input i for the evaluation of DMU e,
y_{rd} is the observed value of output r for DMU d,
x_{id} is the observed value of input i for DMU d,
R is the set of all outputs,
I is the set of all inputs, and
D is the set of all DMUs.

This linear programming model maximizes the sum of weighted outputs of all DMUs rather than just one. The constraints are similar as in Equation 2, with the difference that they now apply for all DMUs rather than for just the one DMU that is being evaluated. As such, the number of constraints increases as a multiple of the number of DMUs in the set. Once populated with input data, the model was computed within the software program LINGO, version 13.0.

Results and Discussion

Because of limitations on the accessibility of safety culture survey data as well as direct measurement data for the proposed model variables, the model was tested using randomized, contrived data. The nature of the cultural measurement instrument suggested for use in this model made it incompatible with the timeframe imposed upon this study. As such results are limited by their ability to inform generalizable conclusions. Yet these remain useful as a means of testing the model as an incremental step toward demonstrating the applicability of DEA as a safety culture benchmarking tool.

Simulated Data and Testing

To test the model in the absence of real experimental data, an artificial set of test data were developed. Testing was completed in three phases. In the first phase, artificial test data were evaluated both through the researcher's algorithm and a step-by-step manual computation based on a Microsoft Excel spreadsheet to ensure proper functioning of the algorithm. Data for five organizations were purposively synthesized with apparent differences in the input and output values, and the results from both the researcher's algorithm and the step-by-step computation were compared for consistency in results, shown in Table 3. In the second phase, test cases of artificial data with known efficiency ranking between

organizations were created and evaluated in the model to ensure the model would provide the correct ranking of efficiency. Finally, testing of the model was completed using random, simulated data. In this way, the model was both verified to ensure computational accuracy and validated to demonstrate that performance was as expected given both purposive and random input selection. A primary limitation to testing of the model is that the literature indicates that no analog measures to the proposed model currently exist. As such, this study takes on some exploratory characteristics. Rather than attempt comparison of model results to an established measure, testing of the model was limited to simulated data and evaluation of whether the results were logically supported by input data. Descriptive statistics of data used in the first phase of testing are presented in Table 2. Excel computation of the DEA problem resulted in the scores for each of the five phase one test organizations as given in Table 3.

Table 2
Test Phase 1 Variable Descriptive Statistics

Variable	Mean	Standard Deviation	Min	Max
% of budget	0.87	0.57	0.35	1.80
% of FTEs	1.41	0.63	0.75	2.30
Hours of annual training	19.40	7.47	12.00	30.00
Top leadership message score	1.48	0.43	1.00	2.10
Safety score	4.32	1.37	2.80	6.30

Table 3
DEA Efficiency Estimates for Phase 1 Testing

Organization	DEA Estimate
A	1.00
B	1.00
C	0.98
D	1.00
E	1.00

The test data were evaluated in the researcher's algorithm for the DEA model. The results from the authors' algorithm were consistent with those from the step-by-step manual computation, providing an initial verification of the accuracy of the researcher's algorithm. For second phase testing, artificial data for test DMUs were created. Table 4 shows the artificial test data. Table 5 shows the relative

efficiency expected from the DEA for this test set of data, as well as the efficiencies computed from DEA. This test verified that for predictable situations, that researcher's algorithm provides accurate, logical efficiency rankings.

The third phase, randomized test dataset included data for the four input variables and the one output variable for 14 organizations. The number of test organizations was limited to 14 due to an a priori testing constraint limit of 250, and 14 organizations with 4 inputs and 1 output resulted in 225 constraints. The input data were generated by choosing at random from a normal distribution, with lower-bound truncation at zero for values that fell below zero for all variables, upper-bound truncation at 75 for percentage input values, and upper-bound truncation for top leadership message at 4. The parameters of the normal distribution are shown in Table 6.

Table 4
Phase 2 Test Data

Inputs					Output
Org.	% Budget	% FTEs	Training Hours	Leadership Messages	Safety Culture
A	1	1	1	1	7
B	0	0	15	0	3
C	0	0	30	0	3
D	0	0	60	0	6
E	0	0	45	0	3

For each of the 14 test organizations, a value for the output variable safety culture score was generated based on the four input variables of each organization and a random noise. Equation 4 describes the formula used to generate the safety culture score.

$$\begin{aligned}[\text{safety culture}] = \; & 7 - 7 / (1 \\ & + .6 * [\%\ \text{of budget}] \\ & + .4 * [\%\ \text{of FTEs}] \\ & + .02\ [\text{training hours}] \\ & + .4\ [\text{leadership message}]) \\ & + \varepsilon \end{aligned} \tag{4}$$

Table 5
Expected and Actual Efficiency Ranking

Expected			DEA evaluation	
Org.	Rank	Rationale	Efficiency	Rank
A	1	Highest possible outcome with lowest possible inputs ensures this organization lies on the frontier	1.00	1
B	1	Highest possible outcome to input ratio with only training input, ensures this organization lies on the frontier as input mix different than A	1.00	1
C	3	Second highest possible outcome to input ratio with only training input	.50	3
D	3	Linear multiple (2x) of C for both inputs and outputs	.50	3
E	5	Lowest possible outcome to input ratio with only training input	.33	5

Table 6
Phase 3 Input Variable Parameters

Input variable	Mean	Standard Deviation
% of budget	1.0	.5
% of FTEs	2	.75
Hours of annual training	32	6
Top leadership message score	1.75	.5

This equation was developed to ensure that safety culture increases with increasing values for the input variable and that the mean safety culture score was 5.1, consistent with prior research validating the selected measurement instrument (Wiegmann, von Thaden, Mitchell, Sharma, & Zhang, 2003). Each input variable value range was developed by researcher consensus based on anecdotal experience. The random noise ε, which is randomly drawn from a normal distribution with zero mean and standard deviation 0.3, ensured that efficiencies varied between different organizations. To ensure the safety score values were valid on the scale of the variable, the values were truncated at a lower bound of 1 and an upper bound of 7. Table 7 shows the descriptive statistics of the test data. The full test data is reproduced in Table 8. The safety score mean of 4.83 is about .5 standard deviations

lower than the design target of 5.1. This difference is due to sampling, and these data were found to be suitable for testing of the model.

Figure 1 shows a plot of the relative technical efficiency scores compared to safety culture for the 14 test organizations. It is important to note that some of the relatively most efficient organizations have relatively low safety culture scores. Organization G has the overall lowest safety culture score of 3.93. However, it is also among the efficient organizations. This shows that organization G is able to efficiently use very few inputs and still achieve a safety culture score of 3.93.

Table 7
Phase 3 Test Variable Descriptive Statistics

Variable	Mean	Standard Deviation	Min	Max
% of budget	1.19	0.50	0.30	1.98
% of FTEs	1.79	0.50	0.88	2.40
Hours of annual training	23.32	5.72	13.52	32.48
Top leadership message score	1.61	0.39	1.11	2.23
Safety score	5.13	0.71	3.93	6.21

Table 8
Phase 3 Test Data

	Inputs				Output	
Org.	% budget	% FTEs	Trng. hours	Leadership msg.	Safety Culture	Efficiency
A	0.81	2.00	24.69	2.09	5.51	0.76
B	1.27	2.18	28.54	1.35	4.35	0.68
C	1.57	1.49	21.11	2.09	5.37	0.87
D	1.76	1.97	25.20	1.83	5.97	0.84
E	1.71	1.31	24.36	1.40	5.83	1.00
F	1.98	2.40	19.67	1.44	4.37	0.74
G	0.30	0.89	13.78	1.18	3.93	1.00
H	0.50	2.26	19.35	1.11	4.53	1.00
I	1.32	2.18	13.52	1.57	5.67	1.00
J	1.13	2.05	31.54	1.16	6.21	1.00
K	1.14	0.88	26.66	1.32	4.76	1.00
L	0.76	1.36	32.48	2.04	5.25	0.84
M	0.80	2.08	24.69	2.23	5.49	0.73
N	1.55	1.96	20.94	1.68	4.54	0.71

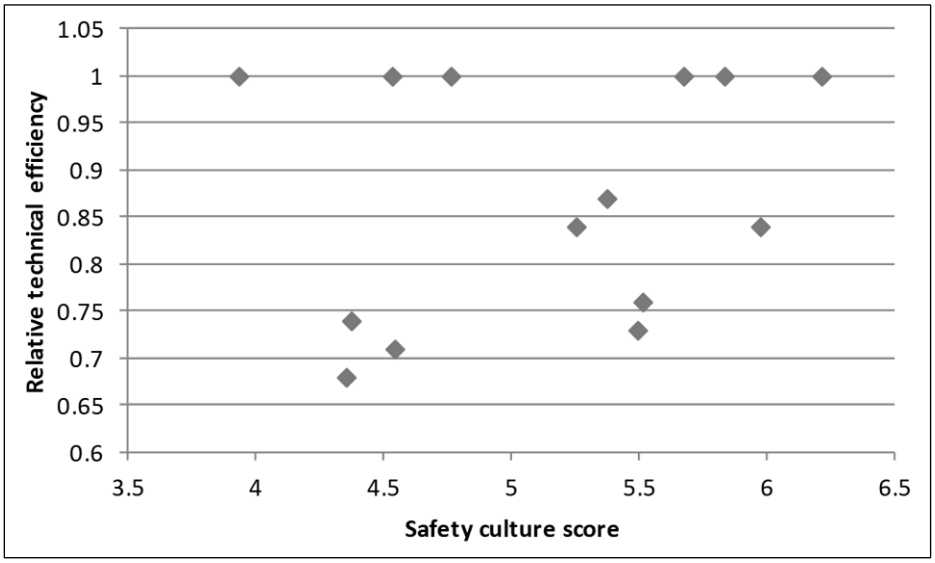

Figure 1. Relative technical efficiency scores versus safety culture.

Bootstrapped DEA

As discussed earlier, one of the primary limitations of DEA is that it is an inherently deterministic model, and it does not make allowances for measurement error in the derivation of efficiency scores. As an extension of the technique previously outlined, bootstrapping allows for the flexibility of DEA to be maintained while allowing for statistical properties of efficiency estimates to be determined. Bootstrapping in the context of DEA was introduced by Simar and Wilson in 1998 based on earlier work introducing the bootstrap by Efron in 1979; and it has since been refined to represent an established technique, as is described in their 2007 article. At its core, bootstrapping is a technique based on the idea of repeated simulation of the data-generating process (DGP) (Simar & Wilson, 1998). Bootstrapping repeatedly resamples the original estimator in order to mimic the sampling distribution missing from deterministic methods. Because the asymptotic distribution in DEA is not only difficult to determine, but also prone to error, bootstrapping is used to estimate the sample properties without the need for fully specifying the DGP (the sample serves as the population from which estimates are treated as valid samples). In their seminal 2007 work, Simar and Wilson describe the advantages of their bootstrapping strategy and test it through a series of Monte Carlo simulations. Simar and Wilson (2007) address concerns regarding the sensitivity of nonparametric frontier approach to outliers and to the inclusion of

stochastic noise, both of which are efficiently suppressed by the double-bootstrapping technique (Murillo-Zamorano, 2004; Simar & Wilson, 2007). In short, bootstrapping allows for statistical inferences to be made within the domain of DEA, a traditionally nonparametric technique.

This study suggests that the concept of performance may be extended to those measures that make up a composite evaluation of safety culture. The literature supports that by using a safety culture assessment instrument to gain performance inputs, DEA can be applied to form a mathematical model that estimates the efficiency of organizational safety culture with respect to peer organizations. By mitigating some of the limitations of traditional DEA and allowing for stochastic analysis, bootstrapped DEA is an appropriate and feasible methodological approach to the issue of benchmarking safety culture in aviation organizations taking into consideration the compounding effects of environmental determinants of efficiency. Further discussion along these lines is included in the results and conclusions that follow.

Stochastic Model

As an expansion of the previously discussed native DEA model, a stochastic form of the same is presented here following the methods outlined by Simar and Wilson (2007). To allow bootstrapping, an output-oriented DEA efficiency estimator must first be calculated as was illustrated by the linear programming model in equation 3. The efficiency estimator hereafter referred to as $\hat{\delta}_i$, can now be used in a two-stage bootstrapping procedure. Even a cursory review of the relevant literature illustrates that in many cases of bootstrapping, a single bootstrap is utilized in conjunction with ordinary least squares (OLS) regression of DEA estimates on covariates treated as environmental variables (Simar & Wilson, 2007). Equation 5 illustrates a common model by which this is accomplished.

$$\hat{\delta}_i = z_i\beta + \varepsilon_i \tag{5}$$

This method presents a problem in that serial correlation is almost always present in DEA efficiency scores, given that calculating an efficiency estimate of one DMU necessarily requires an evaluation of all other DMUs in the subject set. This correlation renders normal regression analysis invalid. In addition, environmental variables as previously described are also correlated with the input and output variables used in the model, leading to biased results, especially in small sample sizes (Assaf & Josiassen, 2011). To circumvent these issues, the double bootstrap proves well suited. As outlined by Simar and Wilson (2007) and as in

Coelli, Perelman, and Romano (1999), Equation 6 represents a second-stage estimation to determine what factors contribute to variation in production efficiency. Size and accident rate for each DMU are proposed as appropriate environmental actors for second-stage estimation and analysis.

$$\hat{\delta}_{it} = \beta_0 + \beta_1 Size_{it} + \beta_2 Rate_{it} + \varepsilon_{it} \quad (6)$$

Where: $\hat{\delta}_{it}$ represents the CCR DEA efficiency estimates,
$Size_{it}$ is a dummy variable wherein organizations of greater than 50 employees, are represented by 1, otherwise, 0,
$Rate_{it}$ is the annual normalized accident rate,
and ε_{it} is random error representing statistical noise as was included in the previously described native DEA model.

This model, once populated with appropriate data, can then be double bootstrapped as described and first- and second-stage estimation values can, assuming that second-stage estimation variables conform to expectations of constant returns to scale, be evaluated by truncated regression (Barros & Dieke, 2008; Simar & Wilson, 2007). In this way, a stochastic evaluation of the effect of the first- and second-stage may be evaluated in terms of both statistical and practical significance.

Practical Implications

In terms of practical implementation, it is intended that the results of the model presented here be used as a tool for identifying standardized practices as well as the most efficient use of limited resources earmarked for safety initiatives designed to improve safety culture. The absence of accidents, for instance, is not an absolute indicator of either effectiveness or efficiency, and the model presented in this study provides a more proactive, granular approach to the evaluation of particular safety-related initiatives in terms of the efficient use of finite resources toward a focus on the most effective interventions. Of interest in the model proposed here is to note to readers that efficiency and effectiveness may exist as separate constructs. Organizations G and J, as examples, are both operating efficiently, though their safety culture total scores, given in Table 4, are entirely different. This disparity is the impetus behind benchmarking of safety and illustrates the need for further empirical investigation of organizational factors contributing to safety.

As a measure of efficiency improvement over time, this model allows an organization to evaluate how different budgetary allocation or cultural intervention

strategies affect actual output in terms of efficiency of use. At present, such evaluation of safety initiatives is generally poorly specified and as previously discussed is in large part based on forensic and imprecise measures such as accident rate. The double bootstrap procedure suggested herein allows for a more robust evaluation of estimation variables, including those related to technical, managerial, and environmental conditions. Reasoning counterfactually about the efficiency of safety investments, as this method allows, is a substantial step forward in not only understanding the role of organizational safety culture but in the purposeful maturation of culture.

Future Research Directions

As was discussed previously, native DEA techniques are particularly susceptible to errors in both measurements and in sampling. It is a deterministic technique, and as such is limited in that native DEA results cannot be extended to hypothesis testing. Instead, stochastic analysis must be achieved using individual input and output measures, and not the DEA results themselves. The results realized in this study are sufficient to inform the development of further models that may address the restrictions inherent to native DEA as a single-solution benchmarking tool. Further research would likely be best focused on the sequential model expansion proposed herein, and on potential means of counterfactual reasoning within such models. For example, the model presented here requires extensive validation of variable selection, a process that must require data unavailable within the scope of the present study. Results from this study could be used as points of comparison for future research as well, and along those lines, the present study would benefit from the availability of actual organization data.

References

Arezes, P. M., & Miguel, A. S. (2003). The role of safety culture in safety performance measurement. *Measuring Business Excellence, 7*(4), 20-28. https://doi.org/10.1108%2F13683040310509287

Assaf, A. G., & Josiassen, A. (2011). The operational performance of UK airlines: 2002-2007. *Journal of Economic Studies, 38*(1), 5-16. https://doi.org/10.1108/01443581111096114

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078-1092. https://doi.org/10.1287/mnsc.30.9.1078

Barros, C. P., & Dieke, P. U. C. (2008). Measuring the economic efficiency of airports: A Simar-Wilson methodology analysis. *Transportation Research Part 4E, 44*(6), 1039-1051. https://doi.org/10.1016/j.tre.2008.01.001

Beriha, G. S., Patnaik, B., & Mahapatra, S. S. (2011). Safety performance evaluation of Indian organizations using data envelopment analysis. *Benchmarking: An International Journal, 18*(2), 197-220. https://doi.org/10.1108/14635771111121676

Beus, J. M., Payne, S. C., Bergman, M. E., & Arthur, W. (2010). Safety climate and injuries: An examination of theoretical and empirical relationships. *Journal of Applied Psychology, 95*(4), 713-727. https://doi.org/10.1037/a0019164

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

Choudry, R. M., Fang, D., & Mohamed, S. (2007). The nature of safety culture: A survey of the state-of-the-art. *Safety Science, 45*(10), 993-1012. https://doi.org/10.1016/j.ssci.2006.09.003

Coelli, T., Perelman, S., & Romano, E. (1999). Accounting for environmental influences in stochastic frontier models: With application to international airlines. *Journal of Productivity Analysis, 11*, 251-273. https://doi.org/10.1023/a:1007794121363

Conklin, T. (2016, September 17). PAPod 88 - Edgar Schein - The father of organizational culture speaks [Audio podcast]. In *Pre-accident* podcast. Retrieved from https://preaccidentpodcast.podbean.com/

Cooper, M. D. (2000). Towards a model of safety culture. *Safety Science, 36*(2), 111-136. https://doi.org/10.1016/s0925-7535(00)00035-7

Cooper, W. W., Seiford, L. M., & Zhu, J. (Eds.). (2011). *Handbook on data envelopment analysis* (2nd ed.). New York, NY: Springer Science. https://doi.org/10.1007/978-1-4419-6151-8

Davies, F., Spencer, R., & Dooley, K. (2001). *Summary guide to safety climate tools* (Offshore Technology Report 1999/063). Norwich, UK: Her Majesty's Stationary Office.

El-Mashaleh, M. S., Rababeh, S. M., & Hyari, K. H. (2010). Utilizing data envelopment analysis to benchmark safety performance of construction contractors. *International Journal of Project Management, 28*(1), 61-67. https://doi.org/10.1016/j.ijproman.2009.04.002

Fang, D. P., Huang, X. Y., & Hinze, J. (2004). Benchmarking studies on construction safety management in China. *Journal of Construction Safety Engineering and Management, 130*(3), 424-432. https://doi.org/10.1061/(asce)0733-9364(2004)130:3(424)

Federal Aviation Administration (2015, January 8). *Safety management systems for aviation service providers.* (AC120-92B). Washington, DC: Author. https://doi.org/10.4324/9781315607504

Flight Safety Foundation (n.d.). Ground Accident Prevention (GAP). Retrieved September 19, 2011, from http://flightsafety.org/archives-and-resources/ground-accident-prevention-gap

Flin, R., Mearns, K., O'Connor, P., & Bryden, R. (2000). Measuring safety climate: Identifying the common features. *Safety Science, 34*(1-3), 177-192. https://doi.org/10.1016/s0925-7535(00)00012-6

Frost, P. J., Moore, L. F., Louis, M. R., Lundberg, C. C., & Martin, J. (Eds.). (1991). *Reframing organizational culture.* Newbury Park, CA: SAGE Publications. https://doi.org/10.1002/job.4030160510

Fuller, C. W. (1997). Key performance indicators for benchmarking health and safety management in intra- and inter-company comparisons. *Benchmarking for Quality Management & Technology, 4*(3), 165-174. https://doi.org/10.1108/14635779710181406

Geller, E. S. (1994). Ten principles for achieving a total safety culture. *Professional Safety, 39*(9), 18-24.

Guldenmund, F. W. (2010). (Mis) understanding safety culture and its relationship to safety management. *Risk analysis, 30*(10), 1466-1480. https://doi.org/10.1111/j.1539-6924.2010.01452.x

Ingalls Jr., T. S. (1999). Using scorecards to measure safety performance. *Professional Safety, 44*(12), 23-28.

International Civil Aviation Organization (2013). *Safety management manual: Third edition* (Doc 9859 AN/474). Montreal, Canada: Author.

Jenkins, L., & Anderson, M. (2003). A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European Journal of Operational Research, 147*(1), 51-61. https://doi.org/10.1016/s0377-2217(02)00243-6

Keren, N., Mills, T. R., Freeman, S. A., & Shelley, M. C. (2009). Can level of safety climate predict level of orientation toward safety in a decision making task? *Safety Science, 47*(10), 1312-1323. https://doi.org/10.1016/j.ssci.2009.01.009

McSween, T. E. (2003). *The values-based safety process* (2nd ed.). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471721611

Murillo-Zamorano, L. R. (2004). Economic efficiency and frontier techniques. *Journal of Economic Surveys, 18*(1), 33-77. https://doi.org/10.1111/j.1467-6419.2004.00215.x

O'Connor, P., Buttrey, S. E., O'Dea, A., & Kennedy, Q. (2011). Identifying and addressing the limitations of safety climate surveys. *42*(2), 259-265. https://doi.org/10.1016/j.jsr.2011.06.005

O'Connor, P., O'Dea, A., Kennedy, Q., & Buttrey, S. E. (2011). Measuring safety climate in aviation: A review and recommendations for the future. *Safety Science, 49*(2), 128-138. https://doi.org/10.1016/j.ssci.2010.10.001

Peterson, D. (2001). The safety scorecard: Using multiple measures to judge safety system effectiveness. *Occupational Hazards, 63*(5), 54-58.

Pidgeon, N. (1998). Safety culture: Key theoretical issues. *Work & Stress, 12*(3), 202-216. https://doi.org/10.1080/02678379808256862

Ray, S. C. (2004). *Data envelopment analysis: Theory and techniques for economics and operations research.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511606731

Ray, S. C. (2005). Input aggregation in models of data envelopment analysis: A statistical test with an application to Indian manufacturing. *Economics Working Papers.* Paper 200554

Reason, J. (1997). *Managing the risks of organizational accidents.* Aldershot, UK: Ashgate Publishing Limited. https://doi.org/10.4324/9781315543543

Schein, E. H. (2010). *Organizational culture and leadership* (Vol. 2). John Wiley & Sons.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science, 44*(1), 49-61. https://doi.org/10.1287/mnsc.44.1.49

Simar, L., & Wilson, P. W. (2001). Testing restrictions in nonparametric efficiency models. *Communications in Statistics-Simulation and Computation, 30*, 159-184. https://doi.org/10.1081/sac-100001865

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics, 136*(1), 31-64. https://doi.org/10.1016/j.jeconom.2005.07.009

Teo, E. A., & Ling, F. Y. (2006). Developing a model to measure the effectiveness of safety management systems of construction sites. *Building and Environment, 41*(11), 1584-1592. https://doi.org/10.1016/j.buildenv.2005.06.005

von Thaden, T. L., Wiegmann, D. A., Mitchell, A. A., Sharma, G., & Zhang, H. (2003). Safety culture in a regional airline: results from a commercial aviation safety survey. In *12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University.

Wiegmann, D. A., von Thaden, T. L., Mitchell, A. A., Sharma, G., & Zhang, H. (2003, February). *Development and initial validation of a safety culture survey for commercial aviation* (AHFD-03-3/FAA-03-1). Washington, D.C.: Federal Aviation Administration. https://doi.org/10.1207/s15327108ijap1602_6

Wiegmann, D. A., Zhang, H., von Thaden, T. L., Sharma, G., & Mitchell Gibbons, A. (2004). Safety culture: An integrative review. *International Journal of Aviation Psychology, 14*(2), 117-134. https://doi.org/10.1207/s15327108ijap1402_1

Zohar, D. (1980). Safety climate in industrial organizations: Theoretical and applied implications. *Journal of Applied Psychology, 65*(1), 96-102. https://doi.org/10.1037//0021-9010.65.1.96