IIE Transactions (2011) **43**, 220–230

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2010.491497

Performance measurement in the warehousing industry

ANDREW JOHNSON[1,*] and LEON McGINNIS[2]

[1]*Department of Industrial and Systems Engineering, Texas A&M University, 237K Zachry Engineering Center, College Station, TX 77843, USA*
E-mail: ajohnson@tamu.edu
[2]*Keck Virtual Factory Laboratory, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, GA 30332, USA*
E-mail: leon.mcginnis@isye.gatech.edu

Received February 2008 and accepted April 2010

Warehouses are a substantial component of logistic operations and an important contributor to speed and cost in supply chains. While there are widely accepted benchmarks for individual warehouse functions such as order picking, little is known about the overall technical efficiency of warehouses. Lacking a general understanding of warehouse technical efficiency and the associated causal factors limits industry's ability to identify the best opportunities for improving warehouse performance. The problem is compounded by the significant gap in the education and training of the industry's professionals. This article addresses this gap by describing both a new methodology for assessing warehouse technical efficiency based on empirical data integrating several statistical approaches and the new results derived from applying the method to a large sample of warehouses. The self-reported nature of attributes and performance data makes the use of statistical methods for rectifying data, validating models, and identifying key factors affecting efficient performance particularly appropriate. This article also identifies several opportunities for additional research on warehouse assessment and optimization.

[Supplementary materials are available for this article. Go to the publisher's online edition of *IIE Transactions* for appendices and additional tables.]

Keywords: Warehouse, facility logistics, data envelopment analysis, outlier detection, two-stage DEA

1. Introduction

Performance assessments in warehousing are needed to identify the options in design and operations that confer the greatest benefits (i.e., "speeding up" the supply chain, minimizing order picking costs, etc.). There are two related but distinct approaches to performance measurement: economic (i.e., revenue related to cost) and technical (i.e., outputs related to inputs). Economic performance assessment is somewhat difficult because warehouses typically do not generate revenues; rather, their function is to support the supply chain including bricks-and-mortar and web-based outlets. Moreover, since a firm's warehouses can be sited in urban, rural, or international locales, the differences in the settings will have a major impact on the costs of the resources used by each warehouse, such as labor and building space. Furthermore, the acquisition costs of capital equipment specific to warehouses vary depending on general economic conditions and the buying power of the specific warehouse owner (e.g., large 3PL versus start-up company). For these and other reasons, technical measures based on output generated and resources consumed tend to give a clearer picture of operational performance when assessing warehouses across a group of warehouses because the measures avoid the uncertainty or variation introduced when using financial measures directly.

Technical performance measurement in the warehouse industry traditionally employs a set of single-factor productivity measures that compare one output to one resource (or input). This is sometimes called the *ratio method* (see Tompkins *et al*. (2003) and Chen and McGinnis (2007)). However, using a set of ratio measures can lead to confusion—if some measures are good and some are poor, is the warehouse performing well? Thus, it is more useful to employ a measure that considers simultaneously all of the significant inputs and outputs.

The field of *production economics* (Coelli *et al*., 2005; Fried *et al*., 2007) provides a variety of approaches to the assessment of technical efficiency when there are multiple inputs and outputs. This article uses the approach of Data Envelopment Analysis (DEA; Charnes *et al*. (1978)) and presents several adaptations that make the approach more applicable to self-reported warehouse data and summarizes

*Corresponding author

the results of applying the adaptations to a large and diverse sample of warehouses. Data on warehouse performance collected over a 5-year period (2001–2005) are used to benchmark the performance of each observation against all other observations in the data set. The purpose of the reported study is twofold: (i) to develop useful methods by which both individual warehouses and groups of warehouses can be evaluated with regard to technical efficiency; and (ii) to identify the operational policies, design characteristics, and attributes of warehouses that are correlated with greater technical efficiency. The results reported are reflective of the data set and methods used. To the extent that the data are representative of general warehouse operations and the practitioner accepts the assumptions related to the models and methods used, the conclusions of this article are reflective of warehousing best practices for general warehousing operations.

2. Literature review

Surprisingly, the technical literature on warehouse performance assessment is meager. Two streams can be identified: papers that propose a framework for designing or analyzing warehouses and those that directly address performance assessment. Rouwenhorst *et al.* (2000) and Gu *et al.* (2007) typify the first category. Both address the coordination problems that arise from the investigation of warehousing subproblems. Rouwenhorst *et al.* (2000) suggest a framework in which to place these problems, but it is largely descriptive and does not provide an operational technique to coordinate the design decisions. Gu *et al.* (2007) categorize the decision problems associated with design and operation rather than overall warehouse performance assessment.

The second category, logistics benchmarking, is used by more than half of the Fortune 1000 companies to improve productivity and quality (Foster, 1992). However, prior to the start of the benchmarking project discussed in this paper (in 2001), relatively few warehouse benchmarking results are described in the literature. Three notable exceptions are Stank *et al.* (1994), Cohen *et al.* (1997), and Hackman *et al.* (2001). Stank *et al.* (1994) gathered survey data from 154 warehousing companies to determine if they employed benchmarking and in what specific areas. The authors then looked at size and services offered to find correlations with benchmarking practices and operations. In contrast, Cohen *et al.* (1997) used a variety of performance metrics to evaluate service parts warehouses, but their results are somewhat confusing when the authors describe the relationship between performance and inventory practices. However, they do address the service parts system as a whole.

More recent work includes Collins *et al.* (2006), which described the collection of warehouse metrics; i.e., picking and inventory accuracy, storage speed, and order cycle time, that are used in a multi-attribute utility theory analysis to determine the best-performing warehouses in a group of 14. De Koster and Warffemius (2005) performed an international comparison across a set of 65 warehouses operating in Asia, Europe, and the United States in 2000 that used various performance measures to identify differences in performance between warehouses on different continents and warehouses operated by third-party logistic providers or self-operated warehouses. The authors concluded that performance across countries and operating parties is very similar. De Koster and Balk (2008) updated De Koster and Warffemius (2005) by gathering data in 2004 on 39 of the 65 warehouses in the previous study and analyzed the two sets of warehouses using DEA. Their key finding was that European warehouses, which are largely third-party providers, are more efficient than Asian or American warehouses.

Hackman *et al.* (2001) developed a model of a warehouse system using labor, space, and investment as resources, and broken case lines, full case lines, pallet lines, accumulation and a derived quantity they termed "storage function" as services produced. This model was designed to answer three questions:

(a) Do larger warehouses perform more efficiently?
(b) Do capital-intensive warehouses perform more efficiently?
(c) Do non-union facilities outperform their union counterparts?

Data collected for 57 warehouses operating between 1992 and 1996 were analyzed using a DEA model to quantify efficiency. Hackman *et al.* (2001) concluded that smaller, less capital-intensive warehouses are more efficient, and that unionization does not appear to impact efficiency.

The research reported in this article addresses the coordination problem by analyzing the warehouse as a single system and quantifying the performance of the warehouse as a whole. Only then are correlations between the operational methods for individual components of the warehouse system and overall warehouse system performance investigated. We extend the model reported in Hackman *et al.* (2001) to test for the statistical significance of each input and output. Identifying the most parsimonious model that captures the general behavior of the warehouse is important because the data requirements for larger models grow rapidly (Simar and Wilson, 2008). While Hackman *et al.* (2001) considered only three variables that may be drivers of inefficiency, two of which are endogenous to the efficiency measure, this article investigates 33 different factors believed to impact warehouse performance. We also introduce a cost-effective data collection methodology that employs self-reporting of data from a very large set of warehouses via the internet (Johnson *et al.*, 2010). Our approach reduces the effect of sample selection bias relative to Hackman *et al.* (2001) and increases the importance of outlier detection.

3. Methods and models used to measure productivity

This section presents the technical details of the assessment model and the methods used to analyze warehouse performance data. The basic methodology, DEA, is described briefly in Section 3.1. A significant issue, particularly for self-reported data, is the detection of outliers—reported data that appear to be either artificial or erroneous. Section 3.2 presents a method for testing model specification and determining the most parsimonious model that explains the data. Section 3.3 describes a two-stage method to identify correlations between efficiency and practices or attributes. Section 3.4 proposes a warehouse model and investigates whether a statistically significant loss of information would occur with a reduction in the model's size. The two-stage method is applied to the new warehousing model to identify correlations between efficiency and three types of properties: (i) operational policies; (ii) design characteristics; and (iii) attributes of the warehouse.

3.1. *DEA efficiency measurement method*

DEA is used widely in applications to measure efficiency (Emrouznejad *et al.*, 2008). However, its application can provide misleading results if not performed with a keen understanding of the underlying economic assumptions. Below we review several of the axioms of DEA to verify the consistency of the analysis method with the warehousing application described in this article.

DEA is a non-parametric efficiency estimation method based on minimal prior assumptions about the production possibility set (Charnes *et al.*, 1978). This is an important characteristic because the warehousing literature does not contain strong hypotheses about the warehousing production function, and in production economics, reliable production function specification tests are not available. In a DEA approach, the set of observed warehouses is used to approximate the Production Possibility Set (PPS). The PPS represents all input and output combinations that actually can be achieved. The boundary of the PPS is called the *efficient frontier* and characterizes how the most efficient warehouses trade off inputs and outputs. DEA constructs a weighted productivity measure:

$$\frac{v_1 y_1 * \ldots * v_S y_S}{u_1 x_1 * \ldots * u_M x_M}, \tag{1}$$

where $\mathbf{x}_i$ is the input usage of input i, $\mathbf{y}_j$ is the output production of output j, and $\mathbf{u}$ and $\mathbf{v}$ are vectors used to aggregate the input and output data. There are M inputs and S outputs. DEA allows each warehouse to determine the vectors $\mathbf{u}$ and $\mathbf{v}$ individually since in practice warehouses may value inputs and outputs differently. These differences may reflect current stocks of inputs or outputs customer base for a particular output or the existence of supply relationships and contracts allowing further inputs to be acquired at prices not available to other warehouses. The minimal convex hull that encompasses the observed warehouse data and maintains the assumptions about the PPS is used to estimate efficiency. Thus, the efficiency estimates from DEA are always optimistic in the sense that a unit receives the benefit of the doubt (Moesen and Cherchye, 1998) and the efficiency estimate is only based on observed warehouse data.

As used in this article DEA encompasses four non-intrusive assumptions:

1. The proper orientation for measuring efficiency can be selected;
2. The assumptions about the PPS hold;
3. The observations give a good representation of the complete production technology;
4. The observations are measured accurately.

The following sections support why the four assumptions are appropriate.

3.1.1. *Orientation for measuring efficiency*

Orientation refers to the direction taken for measuring the distance from a given observation under evaluation to the efficient frontier of the PPS; this distance is a measure of the observation's inefficiency. In the *input orientation* the distance to the efficient frontier is computed as an equiproportional contraction of all inputs, holding outputs constant, that moves the Decision-Making Unit (DMU—in this article, the warehouse) to the efficient frontier. The proportion of input contraction required is the measure of inefficiency. Therefore, if a warehouse argues that it has complete control over the acquisition of inputs but is unable to influence output levels (demand is exogenous), an input orientation may be justified.

Here, the efficiency of warehouse operations is measured from the warehouse manager's perspective and the orientation is determined based on the following rationale. Warehouses are often considered to be cost centers that typically do not generate profit but instead aid in the distribution of goods, preferably at minimal cost. Although the manager makes decisions about operations, typically he/she has little or no control over the outbound flows of goods, which are often influenced by customer demand, advertising, pricing and the like. Therefore, an input orientation is appropriate because the required outputs of the warehouse are defined externally, and it is the manager's responsibility to fulfill the requirements using minimal resources.

3.1.2. *Assumptions about the PPS*

The construction of the PPS is based on several assumptions, some of which are rather weak, yet they have been accepted for decades. Observe a set of n warehouses, call this set $R \equiv \{(\mathbf{X}_j, \mathbf{Y}_j)\}_{j=1}^{n}$ where $\mathbf{X}_j = (x_{1j}, \ldots, x_{mj}, \ldots, x_{Mj})$ is a vector of observed inputs and $\mathbf{Y}_j = (y_{1j}, \ldots, y_{sj}, \ldots, y_{Sj})$ is a vector of observed outputs. Banker *et al.* (1984) summarize these assumptions as follows.

1. Data Envelopment (DE): Given R, a set of n warehouses, and the PPS T, then DE implies $R \subseteq T$

In the warehouse setting the assumption of DE implies that each data point has been measured accurately and represents a warehouse with access to a common set of production technologies.

2. Graph Convexity (GC): $T = \mathrm{co}\,(T)$, with:

$$\mathrm{co}(T) \equiv \{(\lambda\mathbf{X} + (1-\lambda)\mathbf{X}', \lambda\mathbf{Y} + (1-\lambda)\mathbf{Y}') : \times (\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}') \in T, \lambda \in [0, 1]\}$$

for the convex hull.

The assumption of convexity in the input and output space implies that if two warehouses, $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}', \mathbf{Y}')$, are observed to use specified quantities of inputs, X, to generate specific quantities of output, Y, then it is possible for a warehouse to operate using a convex combination of the observed warehouses' input levels, $\lambda\mathbf{X} + (1-\lambda)\mathbf{X}'$, to generate a convex combination of the warehouses' output levels, $\lambda\mathbf{Y} + (1-\lambda)\mathbf{Y}'$. The assumption is often justified on the basis of time divisibility. The inputs specified (i.e., labor, equipment and space) are typically measured on an annual basis. Thus, if a convex combination is suggested, its value can always be achieved only by using the input for some of the year (e.g., leasing a machine for 6 months to achieve 0.5 machines). For a discussion of this standard production economics assumption, see, for example, Varian (1992).

3. Strong Disposability (SD): $T = m\,(T)$, with $m\,(T) \equiv (\omega\mathbf{X}, \eta\mathbf{Y})$ for all $(\mathbf{X}, \mathbf{Y}) \in T$ and $\omega \geq 1$, $\eta \leq 1$ constructing a monotone hull by including input and output pairs using more input than the observed levels of inputs and including pairs producing less than the observed output using the same level of input in the PPS.

SD implies that if a warehouse can meet a given output requirement, with more inputs it still should be able to meet the output requirements.

Applying the minimum extrapolation principle to the maintained assumptions and defining the production possibility set as the conical convex monotone hull of the observations yields:

$$\mathrm{co}(m(R)) \equiv \left\{(x, y) : x \geq \mathbf{X}\lambda,\, y \leq \mathbf{Y}\lambda,\, \sum \lambda = 1, \lambda \in R_+^n\right\}. \tag{2}$$

This implies that the PPS identified is the minimum set of input/output pairs that includes all observed warehouses and is consistent with the maintained assumption that the frontier of the production possibility set is monotonic and convex. These assumptions lend themselves to an implementation using linear programming; however, it is also possible to implement DEA using a least-squares regression approach (Kuosmanen and Johnson, 2010).

3.1.3. *Representativeness of the set of observations*

When calculating efficiency estimates via DEA, the distance from the observation to the efficient frontier is an estimate of the efficiency level. The efficiency level is never certain because the location of the efficient frontier cannot be identified exactly without observing all production possibilities. When only a subset of the production possibilities is observed, the result of DEA can be biased if the sample taken is not random. However, as the number of observations increases, the DEA efficiency estimate converges to the true efficiency values.

3.1.4. *Accuracy of the measurements*

DEA assumes accurate measurement of all data. Since the data set used in this article is self-reported, it is impossible to assure accuracy. Hence, we use an outlier detection method developed by Johnson and McGinnis (2008) to understand the quality of the data set being analyzed and to partially address the need to identify observations that may be measured inaccurately. We adapt the definition of outlier provided by Gunst and Mason (1980, p. 252) "as observations that do not fit in with the pattern of the remaining data points and are not at all typical of the rest of the data." When an outlier is detected, it is analyzed to determine the reason for its identification. A decision is then made to include/exclude the observation from further analysis.

Traditional outlier detection models tend to focus on identifying observations that are over-productive (Wilson, 1995; Simar, 2003). However, in the context of the two-stage method overly inefficient observations can also skew the second-stage results (Chen and Johnson, 2010). The method developed by Johnson and McGinnis (2008) measures efficiency relative to an efficient frontier, constructs an inefficient frontier, and searches for outlier relative to the latter. If a large percentage of observations are flagged with relatively weak criteria, this may indicate that the model defining inputs and outputs is poorly specified or that dissimilar production functions were used within the group of observations. The use of an outlier detection method removes dissimilar observations. Note that relative to the null hypothesis, all data are comparable, a Type-I error would imply removing data that were comparable, and a Type-II error would imply including data that were not comparable. Observations that are unique from the remaining population of warehouses are identified using a critical level for the super efficiency scores. The critical level is both the percentage decrease in inputs and the percentage decrease in outputs necessary to move an observation under evaluation into the PPS constructed using the other observations in the data set. Larger critical levels increase the likelihood of Type-II errors, while stricter (smaller) critical levels increase the likelihood of Type-I errors. This is a necessary trade-off of any outlier detection method.

The outlier detection process identifies a set of data with high density over which a production function can be estimated with confidence. After identifying the set of data to be analyzed, a model selection test is performed. The sequence of these two tests is arbitrary and there does not appear to be any guidance in the productivity and efficiency literature to indicate the relationship between outlier detection and model specification methods. This is clearly an area for future potential research.

3.2. *The model specification test of Pastor* et al. *(2002)*

Models must be detailed enough to capture the phenomena being investigated but no more detailed than necessary (Varian, 1997). Applying this concept to DEA, Pastor *et al*. (2002) developed a statistical test to identify if an input/output model can be reduced with respect to the total inputs/outputs without a statistically significant loss of information. This is relevant because DEA suffers from the curse of dimensionality, meaning that in order for the DEA efficiency estimates of true efficiency to converge, the data requirements grow exponentially as the number of inputs and outputs increase.

The model of Pastor *et al*. (2002) can be described in three steps. First, a DEA linear program is solved for all observations in the data set for the most detailed model including all inputs and outputs of interest. Second, the same linear programming problem is solved for all observations using a model with fewer inputs and outputs. Since the reduction of inputs and output variables results in fewer constraints in the linear program, the efficiency estimates will either remain unchanged or decrease when inputs or outputs are deleted. Third, a statistical comparison is made of the two distributions of the efficiency estimates for the data set. The results of the third step will measure the impact or information lost by reducing the model size. If the performance distribution does not change substantially, an insignificant amount of useful information is lost and the reduced model can be used. After specifying, a production model efficiency can be estimated. There are typically factors that are believed to influence performance but are not necessarily inputs or outputs, in the sense that these factors may not be determined by the warehouse manager or they may not be substitutable for the other inputs or they may not directly generate the outputs of the production process. These are often referred to as *practice*, *attributes*, or *contextual variables*. A two-stage method will be used to investigate the effects of these variables on efficiency.

3.3. *The two-stage method*

The two-stage method consists in estimating the efficiency in the first stage using DEA and identifying contextual variables correlated with high efficiency in the second stage using a regression model and solving it using ordinary least squares. This technique is traditionally attributed to Ray (1991); however, see also Banker and Natarajan (2008) or Johnson and Kuosmanen (2009) for statistical properties of the estimators.

3.4. *Systems model of warehousing*

Our model measures warehouse performance based on the ratio of the services produced to the resources consumed. It includes the most costly inputs, particularly those that may be substituted for one another (e.g., labor and equipment, equipment, and space), and the most valuable outputs. Considerable work on defining input/output models for warehouses has appeared in Hackman *et al*. (2001). De Koster (2008) described a method for identifying the most important aspect of a warehouse. Furthermore, De Koster and Balk (2008) have presented an insightful discussion regarding variable selection. Building on Hackman *et al*. (2001) and our own discussions with warehouse managers and industry consultants, we develop a set of potential inputs and outputs. Figure 1 shows a possible input/output model for a warehouse.

3.4.1. *Definition of inputs*

Two of the most important inputs are capital and labor (Solow, 1957). For our analysis, we measure labor in hours annually and count both direct and indirect labor hours. Capital is divided into inventory, space, and equipment. Inventory is the on-hand inventory measured in dollars and is an average inventory level for the year. The space component is measured in square feet. Equipment is

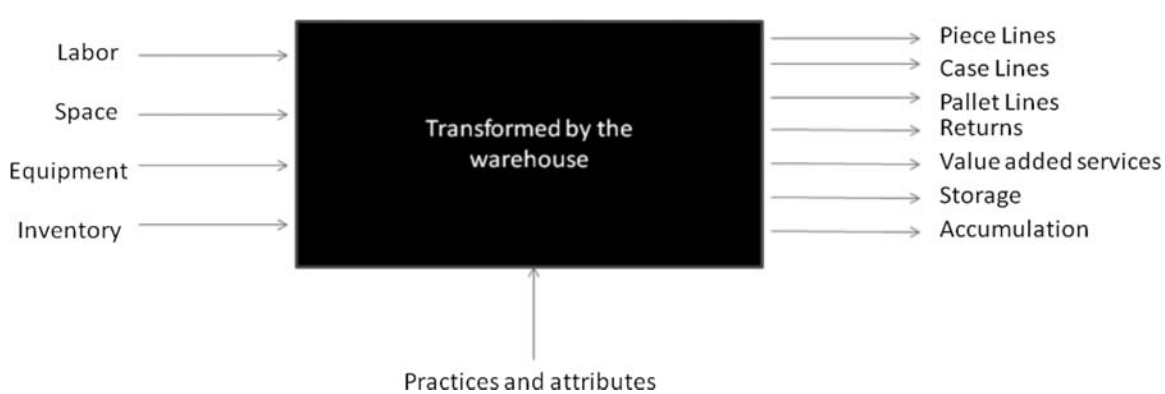

Fig. 1. Proposed input and output variables for a general warehouse.

measured using an equipment inventory as shown in Table A1 in the Appendix online.

To aggregate equipment to a single measure, we apply a cost factor to each equipment category, determining the value from sources such as costing models provided by consultants and equipment vendors. The precise values of these cost factors are less important than their relative values. The resulting aggregate equipment capital cost is a measure used to quantify equipment as an input to the production process in a manner that is consistent across warehouses, industries, and time. For a variety of reasons these aggregate values may differ significantly from the actual price paid by a specific warehouse at a specific point in time. Because we are interested in technical efficiency, the precise value of equipment is not important; rather, a consistent way of comparing equipment inventories is important.

3.4.2. *Definition of outputs*

Outputs are the results of warehousing operations. A warehouse typically exists to fulfill orders and to store products. The orders fulfilled satisfy the downstream customers in the supply chain, and storage is a service provided to the manufacturers. Orders have order lines (or simply, lines), which may require piece picks (piece lines), case picks (case lines), or pallet picks (pallet lines). It is necessary to designate them as different outputs because different types of lines involve different levels of resource commitment. The storage function is a metric developed by Hackman *et al.* (2001) to quantify the capacity of the warehouse to store products while considering the different Stock Keeping Units (SKUs) of product:

$$\pi\sqrt{\text{Number of broken case SKUs}} + (1-\pi) \times (5\sqrt{\text{Number pallet locations}} + \sqrt{\text{floor storage sq. ft.}}) \tag{3}$$

where π denotes the proportion of lines picked as broken cases. Very large orders often require assembly before shipment. The accumulation measure quantifies the difference between the total lines shipped and the total orders that characterize the order assembly effort.

In industries such as publishing, returns are a substantial portion of warehouse labor requirements (Lindsley *et al.*, 1991). Some warehouses are required to change or assemble the products received/stored, usually referred to as Value-Added Services (VAS). Since warehouses with VAS tend to blur the distinction between warehousing and manufacturing systems, we exclude VAS from our model. We also exclude returns processing, because less than 20% of the warehouse data records included this information. Since both VAS and returns are excluded as output measures, the participating warehouse managers were instructed not to include inputs used in these services in their input measures. Not every warehouse will produce non-zero output for each output included in the model, but DEA can still estimate efficiency, because weights for the inputs and outputs are selected for individual warehouses, and a warehouse with zero value for an output can assign it zero weight.

The data set used satisfies the desirable property that all inputs are strictly positive to avoid any infeasibility of the linear program used to calculate efficiency. For details see Zhu (1996) or Johnson and McGinnis (2009).

4. Warehousing data analysis

The data set consists of nearly 400 warehouse records collected over a 5-year period and is treated as a cross section because the technical progress during the time period is believed to be minimal. The data were collected via the iDEAs-W web site (http://ise.tamu.edu/ideas). iDEAs-W provides a browser-based interface that allows users to enter data and receive an efficiency estimate based on the warehouse data collected to that point. It has been recognized by both industry and academia as a pioneering benchmarking tool (Anon, 2002; Dyckman, 2001; Hamdan and Rogers, 2008; Johnson et al., 2010; Thanassoulis *et al.*, 2008). Between 2001 and 2003 warehouses using iDEAS-W entered minimal information on practices and attributes. After 2003, more significant information was required; however, users were and are still free to enter minimal information. Each data record summarizes the performance of a warehouse for a 1-year period.

Warehouses that lack data entries for any of the inputs included in the reduced warehouse performance model are excluded. When input/output data exist, but practices and/or attributes do not, the input/output data are used in the first stage, calculating efficiencies. Hence, the number of observations used in the second stage is smaller and varies depending on the factor considered.

4.1. *Outlier detection results*

If observations are included in the analysis that have misreported data or are not measured correctly this could be detrimental to DEA, which is deterministic and requires all observations be measured exactly. Because this study uses online self-reported data, this issue is of particular concern.

The outlier detection method of Johnson and McGinnis (2009) is used with a very loose critical level of 1.5, corresponding to a 50% increase in inputs, or using the inverse of the critical level, $1/1.5 = 0.66 \Rightarrow$ a 33% decrease in output. If an observation after increasing its inputs is then located within the PPS constructed by the set of all other observations using DEA, it is not flagged as a possible outlier. Although a stricter criterion can also be chosen, we suggest that a loose criterion is appropriate, given the method of data collection. The same critical level is used to identify overly inefficient observations related to an inefficient frontier. After identifying and removing observations that are extremely distant from the data set, 216 observations

Table 1. Descriptive statistics for the 390 observations of input and output levels

	Labor (hours)	*Space (sq. ft.)*	*Capital (\$)*	*Broken case lines*	*Full case lines*	*Pallet lines*
Average	213 138	6730 032	1395 844	1974 091	1162 764	70 316
Stand Dev.	579 359	127 663 556	2671 443	9855 129	9462 737	315 111
Minimum	3000	2090	8000	0	0	0
Maximum	9000 000	1920 768	21 000 000	176 758 000	176 758 000	5388 632

remain. Summary statistics characterizing the entire data set and the reduced data set appear in Table 1 and Table 2, respectively. Note that in Table 2 all distributions skew left, indicating many "small" warehouses in the data set. Furthermore, 22% shipped no full case lines and 37% shipped no pallet lines. Additional information describing the differences in the data set before and after the outlier detection process can be found in Tables A2 and A3 in the Appendix online.

4.2. *Specification test results using the model of Pastor* et al. *(2002)*

To apply the model of Pastor *et al.* (2002), a criterion must be specified regarding the percentage of observations impacted and at what level. We follow the recommendation in Pastor *et al*. (2002) that if a variable is removed, the efficiency estimates of warehouse data set should change by less than 10% for 90% of the warehouses. The iterative test consists of the following.

Step 1. Begin with a model and test for reductions in dimensionality.
Step 2. Use the model specification test in Pastor *et al.* (2002) to identify the variable that when removed has the smallest percentage of observations for which the efficiency level changes by more than 10%
Step 3. If the percentage of observations is less than 10%, delete the variable and repeat Step 1. If no variable can be removed, stop.

Table 3 shows the percentage of observations for which the efficiency estimate changes by more than 10% for each step; with each iteration one variable is deleted that impacts less than 10% of the observations until the variables remaining impact more than 10% of the observations.

Using these results we choose a three inputs (labor, space, investment) by three outputs (broken case lines, full case lines, pallet lines) model to estimate the efficiency of each warehouse in the first stage.

4.3. *Individual warehouse benchmark identification and improvement*

DEA formulation under Variable Returns to Scale (VRS) and an input-oriented specification defines an efficiency estimator θ_i^{DEA} for warehouse i as the optimal solution to the following linear programming problem:

$$\theta_i^{\mathrm{DEA}} = \min_{\lambda,\theta} \left\{ \theta \,\middle|\, \mathbf{y}_i \le \sum_{h=1}^{n} \lambda_h \mathbf{y}_h \,; \theta \mathbf{x}_i \ge \sum_{h=1}^{n} \lambda_h \mathbf{x}_h ; \sum_{h=1}^{n} \lambda_h = 1; \lambda_h \ge 0 \;\forall h = 1, \ldots, n \right\}. \tag{4}$$

We use a VRS model with an input orientation that forces the warehouses to benchmark against similarly sized warehouses in terms of output levels. The resulting efficiency estimates allow management to quantify warehouse performance relative to a benchmark warehouse that is a convex combination of the observed warehouses. Multipliers λ_i are referred to as intensity weights (used for constructing convex combinations of the observed firms). Given that the benchmark warehouse must also produce at least the same levels of output as the warehouse under evaluation, θ_i^{DEA} quantifies the equiproportional reduction of all inputs a warehouse i should be able to achieve based on the observed behavior of the other warehouses in the data set. The results of Equation (4) allow warehouse i to identify efficient benchmark warehouses as those for which $\lambda_i > 0$. Referring back to the data set allows us to identify operational practices and attributes of the benchmark warehouse. Since multiple warehouses are frequently used

Table 2. Descriptive statistics for the 216 observations of input and output levels

	Labor (hours)	*Space (sq. ft.)*	*Capital (\$)*	*Broken case lines*	*Full case lines*	*Pallet lines*
Average	177 778	207 365	1352 962	1321 126	219 560	45 668
Stand Dev.	108 829	127 174	729 024	961 589	189 043	34 423
Minimum	17 750	25 330	71 500	389	0	0
Maximum	436 000	515 732	2981 000	4170 539	586 837	110 000

Table 3. Variables selected for deletion based on Pastor *et al.* (2002)

	Iteration 1	*Iteration 2*	*Iteration 3*	*Iteration 4*
Services produced				
Accumulation	5	7	7	
Broken case lines	7	8	8	39
Full case lines	18	19	19	27
Pallet lines	25	26	26	32
Storage function	1			
Resources consumed				
Labor	75	76	76	80
Space	57	57	60	61
Inventory	3	3		
Investment	35	36	37	51

to construct the benchmark warehouse, warehouse i will often have multiple sets of operational practices and attributes from which to select. This characterizes the multiple ways by which warehouses can perform efficiently. The benchmark warehouse with the largest λ_i is most similar to warehouse i in the sense that the ratio of input and output levels produced are the most similar. Thus, typically the adaptation of similar operational practices and attributes to the warehouse with the largest λ_i frequently requires the least amount of change. However, in some cases the warehouses on the production frontier may be operating under vastly different conditions. In these cases the warehouse may choose among the others for benchmarking purposes. Methods such as those proposed in Seiford and Zhu (2003) can also be used to identify alternative benchmarks, giving warehouse i a set of recommendations to improve overall performance.

The input-oriented efficiency estimation indicates that 23% of the warehouses operate efficiently, with an average efficiency for the entire sample of 0.66 (or 66%) and a standard deviation of 0.27. This indicates either substantial room for improvement, the existence of attributes that limit performance compared to peers, or both. In Section 4.5 a set of attributes is hypothesized and a second-stage regression is performed to identify the attributes that may hinder performance.

4.4. *Banker's hypothesis test for industry differences*

The 216 warehouses in our analysis operate in a variety of industries; 50 reported their North American Industry Classification (NAIC) code, as shown in Table 4. The warehouses themselves vary; i.e., Maintenance, Repair, and Operating Supplies (MRO); online retailers; etc. Product size ranges from automobile spare parts to compact discs. Our purpose is to measure warehouse technical efficiency and identify associated causal factors that limit the ability of warehouses to perform efficiently and to identify the best opportunities for improving warehouse performance in a general warehousing setting. Thus, we have chosen to analyze the data set collected as a whole. However, it is an interesting research question to see if the data indicate that warehouses in certain industries are at a particular disadvantage relative to warehouses in other industries.

Table 4. NAIC codes, industry descriptions and total warehouses reported

Industry code	*Industry description*	*Total*
42	Merchant Wholesale	1
44	Retail Trade	1
311	Food Manufacturing	3
325	Chemical Manufacturing	1
331	Primary Metal Manufacturing	1
332	Fabricated Metal Product Manufacturing	3
335	Electronic Equipment, Appliance, and Component Manufacturing	2
336	Transportation Equipment Manufacturing	1
420	Wholesale Trader	1
423	Merchant Wholesalers, Durable Goods	5
424	Merchant Wholesalers, Non-Durable Goods	9
425	Wholesale Electronics	2
443	Electronics and Appliance Stores	1
445	Food and Beverage Stores	2
453	Miscellaneous Store Retailers	1
493	Warehousing and Storage	6
511	Publishing Industry	8
811	Repair and Maintenance	2
		50

We use Banker's hypothesis test (Banker, 1993) to determine whether the efficiency distributions of the warehouses in the various industries differ. Table A4 in the Appendix online reports the results. Under the assumption that efficiency follows a half-normal distribution to increase the power of the test over the general Kolmogorov-Smirnov-type non-parametric test, at the 95% confidence level the efficiency distributions of the different industries are indistinguishable, but at the 90% confidence level the Publishing Industry and the Electronics Industry are less efficient than other industries. The Publishing Industry appears significantly less efficient than Merchant Wholesale (42); Fabricated Metal Product Manufacturing (332); Transportation Equipment Manufacturing (336); Merchant Wholesalers, Durable Goods (423); Food and Beverage Stores (445); and Miscellaneous Store Retailers (453). Banker's test also shows that warehouses in Electronic Equipment, Appliance, and Component Manufacturing (335) are far less efficient than Merchant Wholesale (42), Transportation Equipment Manufacturing (336), and Miscellaneous Store Retailers (453).

4.5. *Results of correlation analysis between efficiency and practices or attributes*

Since 2003 iDEAs-W has collected data on a large number of warehouse operational practices and attributes. The list

Table 5. List of warehouse practices and attributes investigated in the second stage for correlation with the efficiency estimates

Number of Replenishments	Average Weight per Order
On-hand Inventory Dollars	Average Cube per Order
On-hand Inventory Units	Use of Warehouse Management Software
Inventory Turns	Performance of Compliant Shipping
Number of SKUs	Average Storage Space Utilization
Pareto Percentage of Items Shipped	Use of Velocity-based Slotting
Pareto Percentage of Inventory Cube	Use of Task Interleaving
SKU Churn	Use of Pick-to-Light
Seasonality	Use of RF Dispatching
Pick Variability	Use of Bar Coding
Planning Horizon	Use of Automated Sortation
Value-added Services	Use of Cross Docking
Response Time	Maintenance Expense (Percentage of Budget)
Multi-story Building	Supervision and Management Expense
Rush Order (Percentage of Orders)	(Percentage of Budget)
Rush Order (Percentage of Lines)	Labor Turnover
Number of Suppliers	Percentage of Temporary Labor

is shown in Table 5. Table 6 shows the factors with the most statistically significant correlations.

We note, however, that factors appearing to be insignificant may be so for several reasons. For example, not all practices/attributes are effective for all warehouses: some may adopt and implement a specific practice/attribute while others do not. This implies that warehouses can adopt a practice/attribute that is inappropriate, hence reducing the apparent effectiveness. Furthermore, these results characterize the 216-observation data set analyzed and depend on the previous method applied in the analysis. To the extent that the observations give a good representation of the complete warehouse production technology, these results are useful to indicate best practices for general warehousing. A brief explanation of how each significant variable affects efficiency is now presented.

Table 6. Practice or attribute factors highly correlated with efficiency

	Observations	*Correlation coefficients*	*Significance level*
Seasonality	40	−0.268	**
SKU churn	41	−0.193	*
SKU Span	29	−0.233	*
Inventory ($)	46	−0.26	*
Total replenishment	43	−0.253	*
Temporary labor	33	−0.413	***
Inventory turns	36	0.342	**
Cross docking	44	0.246	*

*Significant at the 90% confidence level.
**Significant at the 95% confidence level.
***Significant at the 99% confidence level.

Temporary labor (measured as annual hours of temporary labor employed): A temporary worker tends to be less familiar with operations and may need more time to complete a task.

Inventory turns (ratio of a warehouse's annual shipment to its inventory measured in dollars): Rapid turnover requires less storage and thus less space and equipment; because the space is smaller, order pickers travel shorter distances and the warehouse can reduce the levels of all inputs.

Seasonality (volume in the peak month/average volume per month, where volume is based on items): Input levels that fluctuate with seasonal demand make it difficult to adjust space and equipment levels; temporary labor may complicate the scenario; warehouses often support input levels to meet peak period demand at the cost of being less efficient in non-peak times.

Total replenishments (includes the replenishment transactions and is the annual total number of replenishment for all SKUs): Replenishments highly correlate with SKU span (the total number of SKUs in the warehouse) and also with lower efficiency levels; while it is expected when inventory levels can be replenished more often, average inventory levels can be lower and the SKU span effect dominates the lower inventory effect.

Inventory (average inventory level measured in dollars): It can be controlled by the firm through reordering practices. A more efficient warehouse should have better reordering practices to fill orders carrying minimum inventory.

Cross docking (a zero/one variable indicating the manager's response to the question, "Do you perform cross docking?"): Warehouses that use cross docking can often reduce space, equipment, and labor, thus creating more efficiency by eliminating the storage function.

SKU span (total SKUs stored in the warehouse annually): Warehouses with higher complexity and more SKUs often have more difficulty locating particular SKUs and cannot specialize to the same extent as warehouses with fewer SKUs.

SKU churn (percentage of SKUs that change from year to year): Allows a greater variety of products to be supplied to customers; however, often more effort is expended in removing dead stock and slotting new items that do not directly contribute to the outputs identified.

5. Conclusions

The important subject of warehouse performance assessment has been largely ignored in the research literature.

Our analysis has demonstrated the feasibility of coupling DEA with internet-based technologies to empirically assess the technical efficiency for single warehouses and for groups of warehouses. The use of self-reported data required the development of methods for rectifying the data by detecting outliers. The desire to find the most parsimonious model required the application of methods for screening all variables. As with any modeling or analysis approach, the methods applied in this article depend on the underlying assumptions. If these assumptions do not hold, a variety the conclusion drawn may not be valid. However, we assert that DEA, properly applied, is useful for assessing both individual warehouses and groups and lays the foundation for broader, larger-scale deployments of DEA for warehouse benchmarking.

The use of the two-stage method has been shown to be effective for discovering the valid practices/attributes that broadly and significantly impact warehouse efficiency. Some factors are clearly beyond the control of the warehouse or the warehouse manager, implying that any program that benchmarks groups of warehouses (e.g., all warehouses in a firm) should weigh these factors when comparing efficiency results.

Obviously, a more comprehensive empirical study will provide more insights about the factors that affect warehouse technical efficiency, and we suggest a number of directions for additional research and development. The use of technical efficiency should be augmented with the available financial data. It is not yet clear how this could be accomplished, especially in light of the "portfolio" approach to measuring capital input, nonetheless, it is an important issue; e.g., for firms with many warehouses which may each encounter unique local costs. The need to refine models is suggested by our finding that some inputs are not easily adjustable; e.g., investment in automation or size of the warehouse. The use of Johansen's vintage model (Johansen, 1972) or the use of the directional distance function (Chambers *et al.*, 1998) could be helpful in this regard.

It is important to bear in mind that warehouses operate in dynamic environments; a given warehouse will usually operate somewhere on the continuum between a startup situation where there may be considerable excess capacity and a mature situation where throughput has grown to the point that it stresses the available capacity. It is critical that future models and analysis should reflect this reality.

Finally, we suggest that due consideration be given to the relevance for this type of performance assessment to the processes of warehouse design and warehouse improvement. If valid models account for resource inputs and service outputs for warehouses in general, or for warehouses in a particular segment of the industry, these models will be extremely useful in the conceptual design of new warehouses. Further research might examine how to employ such models in the design process.

References

Anon. (2002) How good is warehouse performance? *Modern Materials Handling*, **57**(8), 1.

Banker, R.D. (1993) Maximum-likelihood, consistency and data envelopment analysis—a statistical foundation. *Management Science*, **39**(10), 1265–1273.

Banker, R.D., Charnes, A. and Cooper, W.W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, **30**, 1078–1092.

Banker, R.D. and Natarajan, R. (2008) Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research*, **56**(1), 48–58.

Chambers, R.G., Chang, Y. and Fare, R. (1998) Profit, directional distance function, and Nerlovian efficiency. *Journal of Optimization Theory and Application*, **98**(2), 351–364.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research*, **2**, 429–444.

Chen, W.-C. and Johnson, A.L. (2010) A unified model for detecting efficient and inefficient outliers in data envelopment analysis. *Computers & Operations Research*, **37**(2), 417–425.

Chen, W.-C. and McGinnis, L.F. (2007) Reconciling ratio analysis and DEA as performance assessment tools. *European Journal of Operational Research*, **178**(1), 277–291.

Coelli, T., Rao, D.S.P., O'Donnell, C.J. and Battese, G.E. (2005) *An Introduction to Efficiency and Production Analysis*, Springer, New York, NY.

Cohen, M.A., Zheng, Y.S. and Agrawal, V. (1997) Service parts logistics: a benchmarking analysis. *IIE Transactions*, **29**, 627–639.

Collins, T.R., Rossetti, M.D., Nachtmann, H.L. and Oldham, J.R. (2006) The use of multi-attribute utility theory to determine the overall best-in-class performer in a benchmarking study. *Benchmarking*, **13** (4), 431–446.

De Koster, M.B.M. (2008) Warehouse assessment in a single tour, in *Facility Logistics. Approaches and Solutions to Next Generation Challenges*, Lahmar, M. (ed), Taylor & Francis, Auerbach, NY, pp. 39–60

De Koster, M.B.M. and Balk, B.M. (2008) Benchmarking and monitoring international warehouse operations in Europe, *Production and Operations Management*, **17**(2), 1–10.

De Koster, M.B.M. and Warffemius, P.M.J. (2005) American, Asian and third-party international warehouse operations in Europe: a comparison. *International Journal of Operations and Production Management*, **25**(8), 762–780.

Dyckman, S. (ed) (2001) On-line assessment tool for better warehousing. *The Distributer's & Wholesaler's Advisor*, **13**(18), April 15.

Emrouznejad, A., Parker, B.R. and Tavares, G. (2008) Evaluation of research in efficiency and productivity: a survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences*, **42**(3), 151–157.

Foster, T.A. (1992) Logistics benchmarking: searching for the best. *Distribution*, March, 31–36.

Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (2007) *The Measurement of Productive Efficiency and Productive Growth*, Oxford University Press, New York, NY.

Gu, J.X., Goetschalckx, M. and McGinnis, L.F. (2007) Research on warehouse operation: a comprehensive review. *European Journal of Operational Research*, **177**(1), 1–21.

Gunst, R.F. and Mason, R.L. (1980) *Regression Analysis and Its Application*. Marcel Dekker, New York, NY.

Hackman, S.T., Frazelle, E.H., Griffin, P.M., Griffin, S.O. and Vlasta, D.A. (2001) Benchmarking warehousing and distribution operations: an input-output approach. *Journal of Productivity Analysis*, **16**, 79–100.

Hamdan, A. and Rogers, K.J (2008) Evaluating the efficiency of 3PL logistics operations. *International Journal of Production Economics*, **113**, 235–244.

Johansen, L. (1972) *Production Functions. An Integration of Micro and Macro, Short Run and Long Run Aspects*, North-Holland, Amsterdam, The Netherlands.

Johnson, A., Chen, W.C. and McGinnis, L.F. (2010) Large-scale internet benchmarking: technology and application in warehousing operations. *Computers in Industry*, **61**(3), 280–286.

Johnson, A.L. and Kuosmanen, T. (2009) How do operational conditions and practices effect productive performance? Efficient nonparametric one-stage estimators. Working paper. Available at http://ssrn.com/abstract=1485733 (accessed November 23, 2010).

Johnson, A.L. and McGinnis, L.F. (2008) Outlier detection in two-stage semiparametric DEA models. *European Journal of Operational Research*, **187**(2), 629–635.

Johnson, A.L. and McGinnis, L.F. (2009) The hyperbolic oriented efficiency measure as a solution to infeasibility of super efficiency models. *Journal of the Operational Research Society*, **60**(11), 1511–1517.

Kuosmanen, T. and Johnson, A.L. (2010) Data envelopment analysis as nonparametric least squares regression. *Operations Research*, **58**(1), 149–160.

Lindsley, W.B., Blackburn, J.D. and Elrod, T. (1991) Time and product variety competition in the book distribution industry. *Journal of Operations Management*, **10**(3), 344–362.

Moesen, W. and Cherchye, L. (1998) The macroeconomic performance of nations measurement and perception. Discussion Paper. Departement Economie, Katholieke Universiteit Leuven, Leuven, Belsium.

Pastor, J.T., Ruiz, J.L. and Sirvent, I. (2002) A statistical test for nested radial DEA models. *Operations Research*, **50**(4), 728–735.

Ray, S.C. (1991) Resource-use efficiency in public schools: a study of Connecticut data. *Management Science*, **37**(12), 1620–1628.

Rouwenhorst, B., Reuter, B., Stockrahm, V., van Houtum, G.J., Mantel, R.J. and Zijm, W.H.M. (2000) Warehouse design and control: framework and literature review. *European Journal of Operational Research*, **122**, 515–553.

Seiford, L.M. and Zhu, J. (2003) Context-dependent data envelopment analysis—measuring attractiveness and progress. *Omega*, **31**, 397–408.

Simar, L. (2003) Detecting outliers in frontier models: a simple approach. *Journal of Productivity Analysis*, **20**, 391–424.

Simar, L. and Wilson, P.W. (2008) Statistical inference in nonparametric frontier models: recent developments and perspectives, in *The Measurement of Productive Efficiency and Productivity Growth*, Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (eds), Oxford University Press, New York, NY, pp. 421–521.

Solow, R.M. (1957) Technical change and the aggregate production function. *Review of Economics and Statistics*, **39**(3), 312–320.

Stank, T.P., Rogers, D.S. and Daugherty, P.J. (1994) Benchmarking: applications by third party warehousing firms. *Logistics and Transportation Review,* **30**(1) 55–72.

Thanassoulis, E., Portela, M.C.S. and Despic, O. (2008) Data envelopment analysis: the mathematical programming approach to efficiency analysis, in *The Measurement of Productive Efficiency and Productivity Growth*, Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (eds), Oxford University Press, New York, NY, 251–420.

Tompkins, J.A., White, J.A., *et al*. (2003) *Facilities Planning*. John Wiley & Sons, New York, NY.

Varian, H.R. (1992) *Microeconomic Analysis*, Norton, New York, NY.

Varian, H.R. (1997) How to build an economic model in your spare time, in *Passion and Craft: Economists at Work*, Szenberg, M. (ed), University of Michigan Press, Ann Arbor, MI, pp. 256–271.

Wilson, P.W. (1995) Detecting influential observations in data envelopment analysis. *Journal of Productivity Analysis*, **6**, 27–45.

Zhu, J. (1996) Robustness of the efficient DMUs in data envelopment analysis. *European Journal of Operational Research*, **90**(3), 451–460.

Biographies

Andrew Johnson is an Assistant Professor in the Department of Industrial and Systems Engineering at Texas A&M University. He obtained his B.S. in Industrial and Systems Engineering from Virginia Tech and his M.S. and Ph.D. from the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. His research interests include productivity and efficiency measurement, warehouse design and operations, material handling, and mechanism design. He is a member of the INFORMS, National Eagle Scout Association, and German Club of Virginia Tech.

Leon McGinnis holds the Gwaltney Chair for Manufacturing Systems in the College of Engineering at the Georgia Institute of Technology, with faculty appointments in Industrial and Systems Engineering and Mechanical Engineering. His teaching and research address modeling, analysis, and design of discrete-event logistics systems, and he leads a number of interdisciplinary and industry sponsored research initiatives. He is a member of IEEE and INFORMS and a Fellow of IIE.