



A Benchmarking Framework for Understanding Bus Performance in the U.S.

Journal:	<i>Benchmarking: an International Journal</i>
Manuscript ID	BIJ-08-2019-0367.R1
Manuscript Type:	Original Article
Keywords:	Performance, Benchmarking, Quality, Balanced scorecard, Continuous Improvement, Performance measurement

SCHOLARONE™
Manuscripts

A BENCHMARKING FRAMEWORK FOR UNDERSTANDING BUS PERFORMANCE IN THE U.S.

ABSTRACT

Purpose This paper describes a benchmarking framework applied to medium-sized urban public bus agencies in the United States which has overcome the challenges of data quality, comparability and understanding.

Design/methodology/approach The benchmarking methodology described in this paper is based on lessons learned through seven years of development of a fixed route key performance indicator (KPI) system for the American Bus Benchmarking Group (ABBG). Founded in 2011, the ABBG is a group of public medium-sized urban bus agencies that compare performance and share best practices with peers throughout the United States. The methodology is adapted from the process used within international benchmarking groups facilitated by Imperial College and consists of four main elements: peer selection, KPI system development, processes to achieve high-quality data, and processes to understand relative performance and change.

Findings The four main elements of the ABBG benchmarking methodology consist of eighteen sub-elements, which when applied overcome three main benchmarking challenges; comparability, data quality, and understanding. While serving as examples for the methodology elements, the paper provides specific insights into service characteristics and performance among ABBG agencies.

Research limitations/implications The benchmarking approach described in this paper requires time and commitment and thus is most suitably applied to a concise group of agencies.

Practical implications This methodology provides transit agencies, authorities and benchmarking practitioners a framework for effective benchmarking. It will lead to high-quality comparable data and a strong understanding of the performance context to serve as a basis for organizational changes, whether for policy, planning, operations, stakeholder communication, or program development.

Originality/value The methodology, while consistent with recommendations from literature, is unique in its scale, in-depth validation and analysis, and holistic and multi-dimensional approach.

Keywords: performance, performance management, performance measurement, benchmarking, quality, balanced scorecard, continuous improvement

Research paper

ACKNOWLEDGEMENTS

The authors thank the member agencies of the American Bus Benchmarking Group for their willingness to share their experiences with the development of fixed route benchmarking with the wider transport and benchmarking industry.

INTRODUCTION

Benchmarking performance can be a powerful tool in management. It can provide identification and understanding of strengths and weaknesses, improve processes, and lead to establishing priorities and realistic targets. However, for such benchmarking to be effective, it must be based on high-quality comparable data and a strong understanding of the performance context. The joint development of a fixed-route bus key performance indicator (KPI) system by the members of the American Bus Benchmarking Group (ABBG) and Imperial College London offers an effective methodology for benchmarking bus transit performance. The ABBG's founding agencies wanted to go beyond the scope of the Federal Transit Administration's National Transit Database and adopt a benchmarking methodology used successfully internationally.

Founded in 2011, the ABBG is a group of public medium-sized urban bus agencies that agreed to compare performance and share best practices with peers in similar cities throughout the United States. The ABBG is modeled on the International Bus Benchmarking Group (IBBG), a global group of 16 urban bus systems that was founded in 2004 (Trompet *et al.*, 2009; Trompet *et al.*, 2018).

As of 2020, the ABBG is comprised of 24 agencies, but at the time of this research in 2018, data was available from 19 member agencies, as shown in Figure 1. Performance data from six agencies (indicated with a * and italics) were not yet available at the time of writing, as these organizations joined the ABBG within the past two years. Data was available from WeGo Nashville (indicated with a ^), that was an ABBG member at the time but has since left.

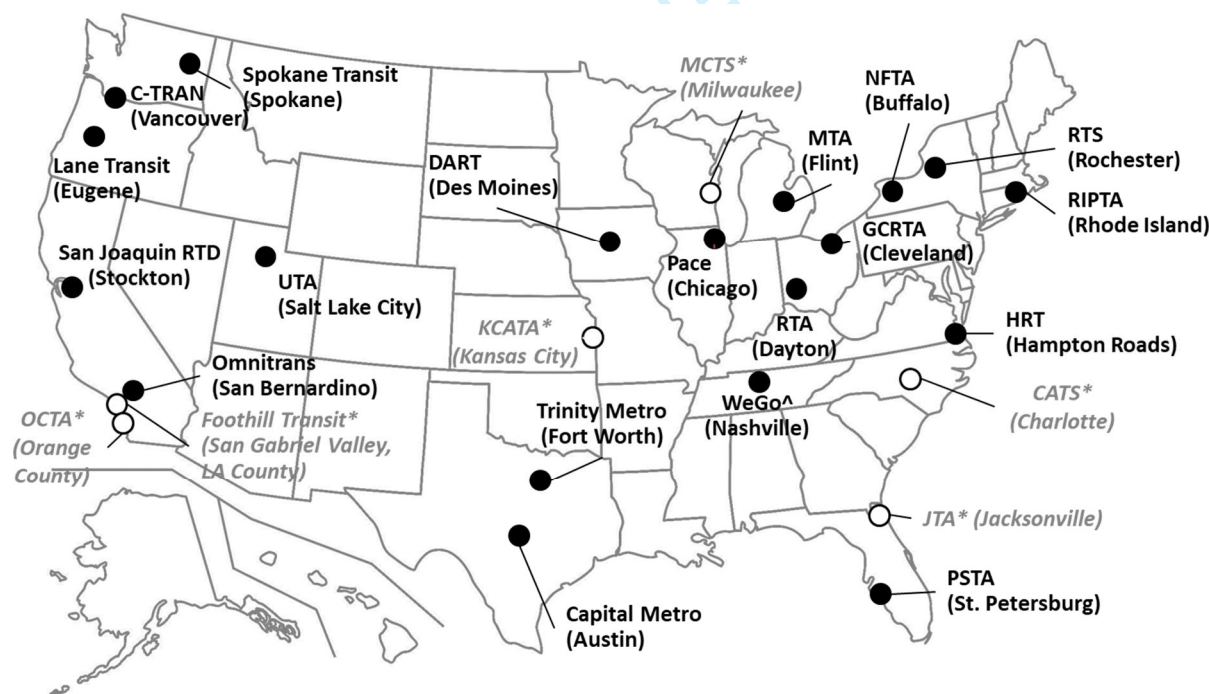


FIGURE 1 Map of ABBG Members (2020) and Research Participants (2018)

Building upon years of international benchmarking experience (especially from the IBBG) and input from ABBG agencies, a fixed-route KPI system was developed from 2011-2014, and dynamically adjusted in the following years to remain in line with agency priorities and technology developments. During the course of development, the following research questions were addressed: what challenges need to be overcome to achieve high-quality comparable data and a strong understanding of the performance context? How can these challenges be addressed in a benchmarking methodology? How can these challenges be addressed within a U.S. context?

Three general categories of challenges were identified that needed to be overcome for the benchmarking system to work well and deliver the desired outcomes (namely, the ability to learn and implement change): comparability, data quality, understanding.

The comparability of data starts with the comparability of the source – in this case, transit agencies – and carries through to comparability of the KPIs selected. Transit agencies can vary in scale, level of service throughout the day and week, management, and types of services. The ABBG’s eligibility criteria and clearly defined scope of benchmarking limit variability and ensure a useful level of similarity while also allowing for benefits from diversity (Trompet *et. al*, 2009). The ABBG’s agency-led prioritization of data and selection of KPIs ensure the benchmarking stays in line with strategic goals.

Achieving high-quality data requires agency commitment, robust definitions and standardized data collection, and in-depth validation processes to ensure accuracy and address differences in data collection and management, which change over time with technology and resources.

Improved data understanding leads to improved understanding of performance. The ABBG approach supports understanding by breaking down data items into subcomponents, using bi/multi-dimensional normalization and multi-KPI analysis, and considering trends and contextual information about the transit agencies’ operating environments and service characteristics.

For the purposes of the U.S. context, specific challenges around different state (and local) contexts, provision of separate paratransit services, and variety of fiscal years also had to be addressed.

For each of the three areas above, this paper describes the challenges and respective solutions identified during the development of the ABBG fixed-route KPI system. To facilitate the discussion, the remainder of the paper is organized as follows: The literature review provides an overview of available relevant academic research and government publications with a focus on U.S. transit performance benchmarking. The methodology section that follows describes the ABBG performance benchmarking process and identifies how elements of the methodology relate to one or more of the three identified challenge areas. The discussion section of the paper then details how these process elements and their respective sub-elements address the three areas of benchmarking challenges listed above, using specific performance and operating characteristic examples from the ABBG. Finally, the paper will provide guidance on how to apply this research, discuss its limitations, and identify opportunities for future work and research.

LITERATURE REVIEW

Performance benchmarking has a rich, robust set of literature, applied to all industries and to all modes of transport (e.g., Hinton *et al.*, 2000). It is worth recognizing that how benchmarking is defined can vary significantly (e.g., as discussed in Alstete, 2008), and that many frameworks

have been identified as successful (e.g., Zairi, 1994 and Geerlings *et al.*, 2006). For public transport specifically, the extensive performance metrics provided by several European Union transport benchmarking projects are often used as a starting point (e.g., EQUIP as covered by Geerlings *et al.*, 2006). Other publications have identified multiple categories of performance indicators such as the eight identified by Kittelson and Associates, Inc. *et al.* (2003) including service delivery, community impacts, travel times, safety and security, maintenance and construction, economics, paratransit, and comfort. Kassens-Noor *et al.* (2019) includes a review of the development of technology-driven performance measures in transit. Perk and Kamp (2004) review various previous attempts to identify composite scores or indexes such as the Transit Performance Index developed by the American Public Transportation Association. Transit service quality and accessibility indices are also common (Fu and Xin, 2007; Al Mamun and Lownes, 2011; Hartgen and Horner, 1997).

The wide body of international literature on transit benchmarking relevant to this research is documented by Trompet *et al.* (2009; 2018). These two publications draw upon this literature as well as the benchmarking experience of the IBBG and provide two key lessons that have been applied to ABBG's methodology. Trompet *et al.* (2009) found benchmarking to be useful and justifiable because there is sufficient variability in performance between agencies due to factors within management control rather than only those resulting from differences in external factors, thus allowing for agencies to learn lessons from one another. Trompet *et al.* (2018) concluded that due to variability of service characteristics between peers, such as commercial speed or trip length, it is necessary that performance be at least reviewed from two different dimensions in order to understand the normalization bias resulting from this variability and to obtain an improved and more balanced understanding of relative performance. These findings are key to the success of the ABBG as described further in this paper.

In Europe, European Union projects such as EQUIP provided both performance metrics and step-by-step approaches (e.g., Geerlings *et al.*, 2006) to follow. In the U.S., there is a well-established body of government-sponsored publications providing both guidance and documentation of fixed route transit performance-measurement systems and peer comparison (Kittelson and Associates, Inc. *et al.*, 2003; Ryus *et al.*, 2010; Florida Department of Transportation, 2014; Perk and Kamp, 2004; Cook *et al.*, 2010). Although the benchmarking methodology described in this paper is consistent with many of the recommendations in these publications, it is distinct in its scale, in-depth validation and analysis, and holistic and multi-dimensional approach. For example, the step-by-step process used in this research shares the cyclical nature and emphasis on learning (and acting upon that learning) of EQUIP as described by Geerlings *et al.* (2006) but is quite different in its starting point. EQUIP begins with a self-assessment rather than a collaborative process to develop common measures. One of the most relevant U.S. benchmarking projects is a recent statewide effort by the Michigan Department of Transportation to measure performance across 80 rural and urban transit systems (Kassens-Noor *et al.*, 2019). Although its goals and performance metrics are quite different, the project also found that its approach also required significant resources and time, trust between participating entities, and one-on-one individual follow-up.

The Federal Transit Administration's National Transit Database (NTD) serves as the main data source for nearly all US public transit benchmarking literature, especially since it was compiled online in an interactive format that allows for peer selection (Gan *et al.*, 2011). The NTD is a valuable large-scale depository of data for the entire U.S. transit industry that began in the 1970s. Submitting data to the NTD is mandatory for nearly all US transit agencies that

receive federal funding and is the basis for formula allocations of federal transit funds. Although fairly comprehensive, the NTD faces challenges in data availability (Verbich *et al.*, 2017) and data quality and comparability, mainly due to its scale, lack of verification, and lack of context and interpretation (Hendren, 2011).

Drawing upon NTD data, many academic publications covering U.S. transit benchmarking focus on algorithms to improve operational efficiency in terms of resource utilization (vehicles and labor) or high-level evaluations of financial efficiency across a large set of agencies (Fielding *et al.*, 1985). One of the well-established quantitative approaches, using NTD data, is data envelopment analysis (DEA). DEA focuses on measuring and comparing efficiency and effectiveness for a very limited number of inputs (e.g., cost, labor, and vehicles) compared to outputs (e.g., boardings, vehicle miles) (e.g., Chu *et al.*, 1992; Karlaftis, 2004). Min and Ahn (2017) applied DEA to over 200 agencies in the US over a multi-year period. Although commendable for adding trends analysis and consideration of important explanatory context (namely organization type, modes, service area density, and residing state), their analysis is limited to one financial and one service efficiency measure.

When benchmarking public transport, the audience is a key consideration. Much of the existing public transport benchmarking literature assumes the policymaker, taxpayer, or transit user (customer) perspective, although Verbich *et al.* (2017) tries to cover all perspectives including the transit agency, Kassens-Noor *et al.* (2019) considers both the transit agencies and the state government, and Hilmola (2011) applies DEA to public transit benchmarking from the city’s perspective. Policymakers are interested in outcomes such as accessibility and affordability (Henning *et al.*, 2011) or environmental emissions and modal share (Hilmola, 2011); taxpayers in cost efficiency and effectiveness (Min and Ahn, 2017); customers in journey time, comfort, and safety (Verbich *et al.*, 2017). The benchmarking methodology described in this paper is primarily intended for the transit agency but attempts to cover performance areas that are priorities for all stakeholders.

METHODOLOGY

Based on a definition by Lema and Price (1995) benchmarking is defined by the ABBG and its sister benchmarking groups as *a systematic process of continuously measuring, comparing and understanding organizations’ performance and change in performance of a diversity of key business processes against comparable peers to gain information which will help the participating organizations to take action to improve their performance*. These principles are repeatedly identified as success factors in benchmarking literature, such as Magd and Curry (2003).

The development of ABBG’s fixed-route benchmarking process and KPI system adopted the methodology taken for the other benchmarking groups facilitated by Imperial College London. Establishing a robust KPI system that uses a comparable, high-quality dataset and obtaining the understanding of underlying differences in performance takes time. While benefiting from lessons learned and experience from the IBBG, it took the ABBG three years of iterative definition development, data collection and analysis before the member agencies were sufficiently satisfied with the level of comparability to be able to use the data for performance comparison. Trompet *et al.* (2009) summarized the necessary benchmarking conditions and decisions necessary in those development years to gradually increase the quality of the dataset.

The methodology that has been the foundation for the establishment, development and continuation of successful benchmarking groups such as the IBBG and applied to the ABBG is presented in Figure 2.

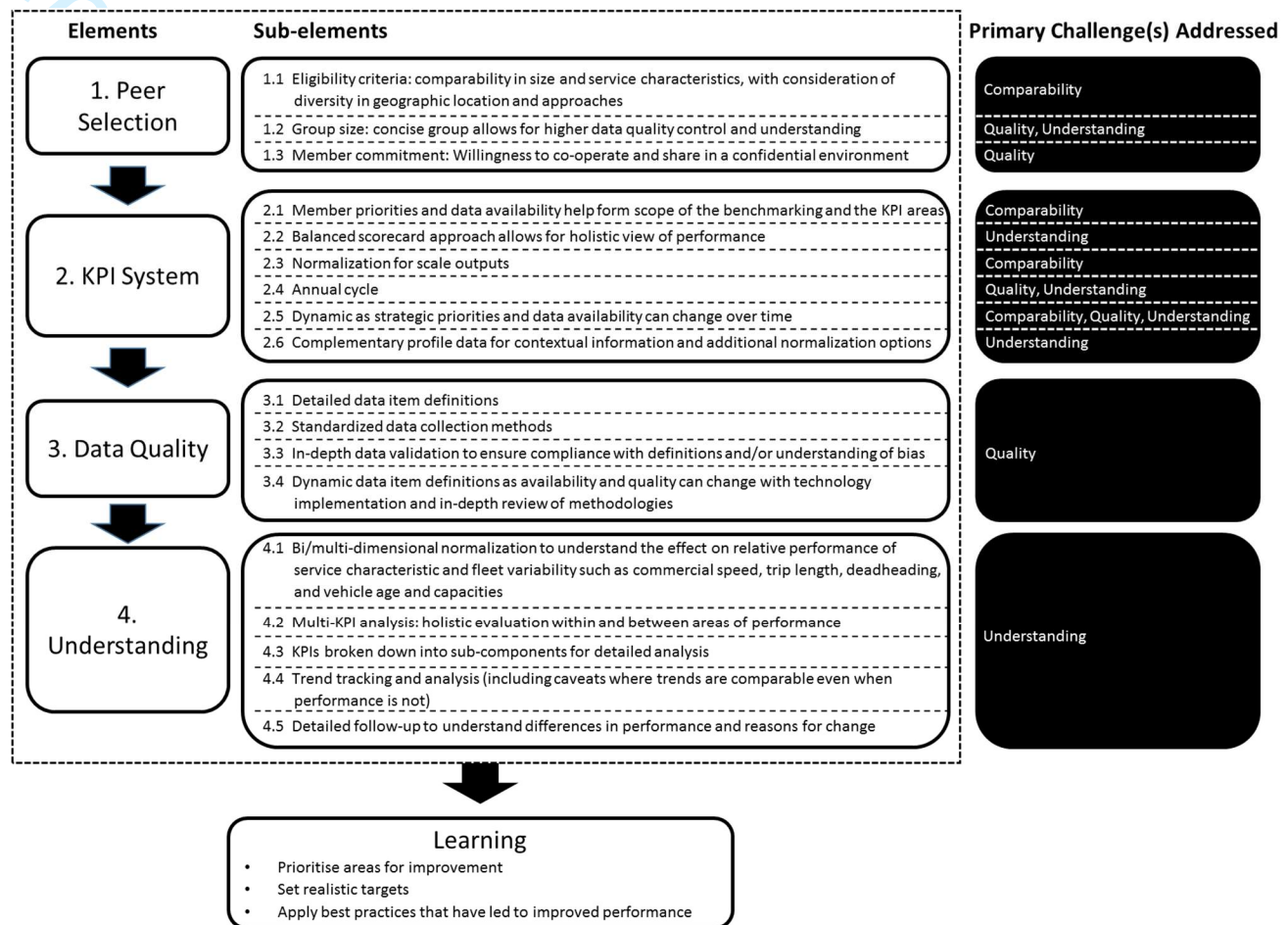


FIGURE 2 ABBG Benchmarking Methodology with Identification of Challenges Addressed by Sub-Element

This methodology consists of four main elements and eighteen sub-elements which, when executed well, enables organizations to learn from the benchmarking and to take action to improve their performance, as expressed by the 'learning' box in the figure. The main elements of the methodology are: peer selection, development of the KPI system, processes to achieve high-quality data, and the analyses and processes to understand relative performance and reasons for change in performance. At the inception of a benchmarking group, these main elements occur in order as expressed through the arrows; however, once the benchmarking group is in the later stages of development and/or established, these main elements occur simultaneously to continuously improve the usefulness and quality of the benchmarking process.

The objective of this paper is to describe how the main elements and respective sub-elements of this benchmarking methodology address the three identified challenges of comparability, data quality, and understanding in benchmarking. Figure 2 outlines how the eighteen sub-elements are linked to one, two, or all three of the benchmarking challenges. The discussion section of this

paper describes these sub-elements and details how each of these methodology elements and sub-elements contribute to addressing the three challenges that needed to be overcome for the benchmarking system to work well and deliver the desired outcomes.

DISCUSSION OF RESULTS

This section is presented in three subsections, each focused on one of the challenges from Figure 2 and the sub-elements (with numbered references given in parentheses) that address the respective challenge.

Challenge: Comparability

This section describes how four of the sub-elements from Figure 2 address comparability.

Eligibility Criteria (1.1)

The ABBG membership criteria, developed by the ABBG founding members, are as follows:

- (1) Mid-sized US or Canadian bus organization,
- (2) Between approximately 5 million and 50 million annual boardings,
- (3) Between approximately 100 and 600 vehicles in fleet,
- (4) Similar service and operating characteristics (fixed-route urban or suburban ‘local’ bus service with all-day, every-day operation with at least 15-30 minute headways on key routes),
- (5) Majority public operation preferred and public management / administration required,
- (6) Willingness to invest resources and fully participate, and
- (7) Geographic diversity.

Generally, the ABBG’s eligibility criteria are designed to balance similar service characteristics (criterion 4) while allowing for variability (criteria 2, 3, and 7) that enables diversity and exchange of new ideas, as documented in Trompet *et al.* (2009). Although there are benefits from benchmarking across multiple countries, the focus on North America (criterion 1) means greater consistency in terminologies, data, operating and regulatory environments, and in the scope of services, as covered below. Differences in scale can be addressed through normalization factors, so the limitations in size (criteria 2 and 3) are due to contextual differences between larger and smaller cities and practical considerations as described under Data Quality. Criteria 5 and 6 are intended to ensure the agency is both able and willing to provide quality data; given that the ABBG was founded to go beyond the NTD, agencies need to have the resources and ability (in terms of technology and data systems) to provide both more and different data than for NTD to meet the definitions agreed upon by the group. The importance of this type of commitment has been documented in benchmarking literature (e.g., Magd and Curry, 2011).

The criteria are applied only to interested parties and then membership is subject to the group’s consensus; the group also has a limitation on size as discussed in the next section.

Scope of Benchmarking (2.1)

To ensure comparability of performance the scope of the ABBG benchmarking effort had to be defined. The group agreed that all regularly scheduled services open to the general public were included in the scope of the ABBG fixed route benchmarking. These include the NTD categories of Motor Bus (MB), Commuter Bus (CB), Trolley Bus (TB), and Bus Rapid Transit (RB), as well as circulators and routes that serve schools or universities but are open to the general public.

In addition, the group agreed consistent with NTD to include route-deviation/flexible services, including those that serve zones, in the fixed route data set. These services are common among the ABBG agencies and considered 'core business'. Excluded services are any special event services or rail replacement.

Unlike in most of the world, all public transit agencies in the U.S. are inherently multimodal because of the requirement for complementary paratransit service, which has very different service characteristics (and thus benchmarking needs) than fixed route service. The ABBG has developed a distinct, separate paratransit benchmarking process and KPI system as described in Morse *et al.* (2017). This separate system complements the fixed route system in terms of both understanding (see below) and comparability, because it accounts for all shared costs and resources across the two data sets and ensures they are allocated between the two modes, especially for those agencies that share drivers and/or maintenance. In addition, seven of the ABBG agencies operate one or more other mode (e.g., light rail, commuter rail, etc.). This has required careful allocation of shared resources (e.g., labor), costs, and revenue.

KPI System (2.1, 2.3, 2.5)

The foundation of the KPI system is a balanced scorecard across four dimensions: Growth & Learning, Internal Processes, Customer, and Financial (Kaplan and Norton, 1992), adapted to transit with the addition of Safety and Environment dimensions as described in Morse *et al.* (2017). Within each success dimension, KPIs were then developed based on agency priorities and data availability or willingness to collect the required data. The fact that the KPIs are agreed by agency senior executives adds to improving the comparability of the KPI system, as KPIs are selected which reflect common interests, challenges, and requirements. Furthermore, the KPI system is dynamic; therefore, the KPI system can be adjusted to reflect changes in strategic goals and/or data availability. For example, it is expected that KPIs that focus on actual customer impact (such as customer delay minutes) will be adopted by the ABBG once technology allows for their measurement. The ABBG fixed-route KPI system at the time of this research (2018) is shown in Table 1. The ABBG also conducts a separate, complementary Customer Satisfaction Survey benchmarking effort (Trompet *et al.*, 2013) as recommended by several studies (e.g., Meyer, 2002).

TABLE 1 ABBG Fixed Route Key Performance Indicator System

Success Dimensions	Key Performance Indicators
Growth & Learning	Ridership: Passenger Boardings (<i>5-year % change</i>) Service Levels: Vehicle Revenue Miles and Hours (<i>5-year % change</i>) Passengers per Revenue Mile & Hour Staff Training
Customer	Customer Information (scheduled and real-time) On-Time Departure Performance Passenger Miles per Revenue Capacity Mile Passenger Miles per Revenue Seat Mile Lost Vehicle Mile Missed Trips
Internal Processes	Peak Fleet Utilization Network Efficiency (revenue miles & hours per total miles & hours, non-revenue split by category) Staff Productivity (total vehicle hours & miles per labor hour, overall and by category) Staff Absenteeism Rate (by staff category and absenteeism type) Fleet Reliability (<i>miles/time between road calls due to technical faults</i>)
Financial	Total Cost per Vehicle Mile & Hour Total Operating Cost per Vehicle Mile and Hour (<i>service operation, maintenance, administration</i>) Service Operation Cost per Revenue Mile & Hour Total Operating Cost per Boarding & Passenger Mile Operating Cost Recovery Fare Revenue per Boarding & Passenger Mile
Safety	Number of Vehicle Collisions per Vehicle Mile and Hour (<i>preventable, non-preventable, on-property</i>) Number of Staff Injuries per Staff Work Hour Number of Passenger Injuries per Boarding & Passenger Mile Number of 3 rd Party Injuries per Vehicle Mile & Hour
Environmental	Fuel Consumption (<i>per total vehicle mile, passenger mile, & capacity mile</i>) Carbon Dioxide Emissions (<i>per total vehicle mile & passenger mile</i>)

Each of the KPIs listed in Table 1 is normalized for scale outputs, which is arguably the sub-element of the system that is most relevant to comparability. The ABBG uses a variety of scale denominators, but there are four primary ones, two on the demand side (passenger boardings and passenger miles) and two on the supply side (vehicle miles and hours, both total and revenue). While most benchmarking exercises and literature focus on the use of mileage as a scale denominator, useful discussions have been provided on the use of vehicle hours as a preferred denominator over vehicle miles (Fielding *et al.*, 1985; Hensher and Daniels, 1995) as much of cost, revenue, and labor are linked to hours rather than miles. The ABBG therefore uses both supply side output denominators.

In addition, specific KPIs were designed to make data more comparable even when there are significant differences between agencies for both service and fleet characteristics. Two examples of such KPIs are passenger miles per revenue capacity mile, a customer utilization KPI that takes into account both the seating and standing capacity of fleets, and carbon dioxide emissions, an environmental KPI that allows for comparison across fleets even with the diversity of fuels used (diesel, diesel hybrid, gasoline, compressed natural gas, electric trolleybus, electric battery bus).

Challenge: Data Quality

This section describes how eight of the sub-elements from Figure 2 address data quality.

Group Principles (1.2, 1.3)

The ABBG has adopted several core principles, including collaboration, depth, speed, and confidentiality. As documented in Trompet *et al.* (2009), one of the main contributing factors to data quality is an open and honest sharing environment that is supported by a strict confidentiality agreement, which ABBG has adopted. However, the group deliberately made the decision to support this publication and share some of its lessons and information with the broader transit industry. Thus, while the graphs and tables in this paper are still anonymized and/or aggregated, they include actual averages, ranges, and in one case (Figure 6), actual KPI values.

These core principles rely on limiting the size of the group to a maximum number that can comfortably meet in person around one table, to facilitate direct discussions and joint decision-making. The limitation to size also makes in-depth validation more manageable. There is also a minimum benchmarking group size required to ensure sufficient comparability within any one KPI, as documented in Trompet *et al.* (2009). From its extensive experience in benchmarking, the research team has determined the ideal size to be somewhere between 8 and 18 agencies. The ABBG has grown beyond this size due to the group's desire to accommodate additional interest and as a consequence, it is in the process of adopting new protocol, including meeting as two subgroups, to ensure the ABBG principles are still met.

Detailed and Dynamic Data Item Definitions (3.1, 3.4)

The ABBG has an advantage compared to international benchmarking in that nearly all U.S. transit agencies are required to submit comprehensive data to the NTD. However, as mentioned previously, the founding ABBG agencies identified a need to go beyond the NTD in terms of quantity and granularity of data and detail of definitions.

To achieve high quality data, all data items necessary to create the KPIs as listed in Table 1 have been defined carefully. Defining and agreeing definitions is a process that requires research and subsequent discussion amongst member agencies. Figure 3 provides an example of how ABBG approaches defining data items. It is a flow diagram showing how the group has agreed to define technical road calls to measure maintenance reliability. All possible aspects of a road call due to technical faults are listed; however, only those elements in black boxes were selected to be included into the ABBG definition. This diagram was developed based on two surveys of ABBG agencies, detailed follow-up with maintenance staff, and lessons learned from a similar process in the IBBG.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

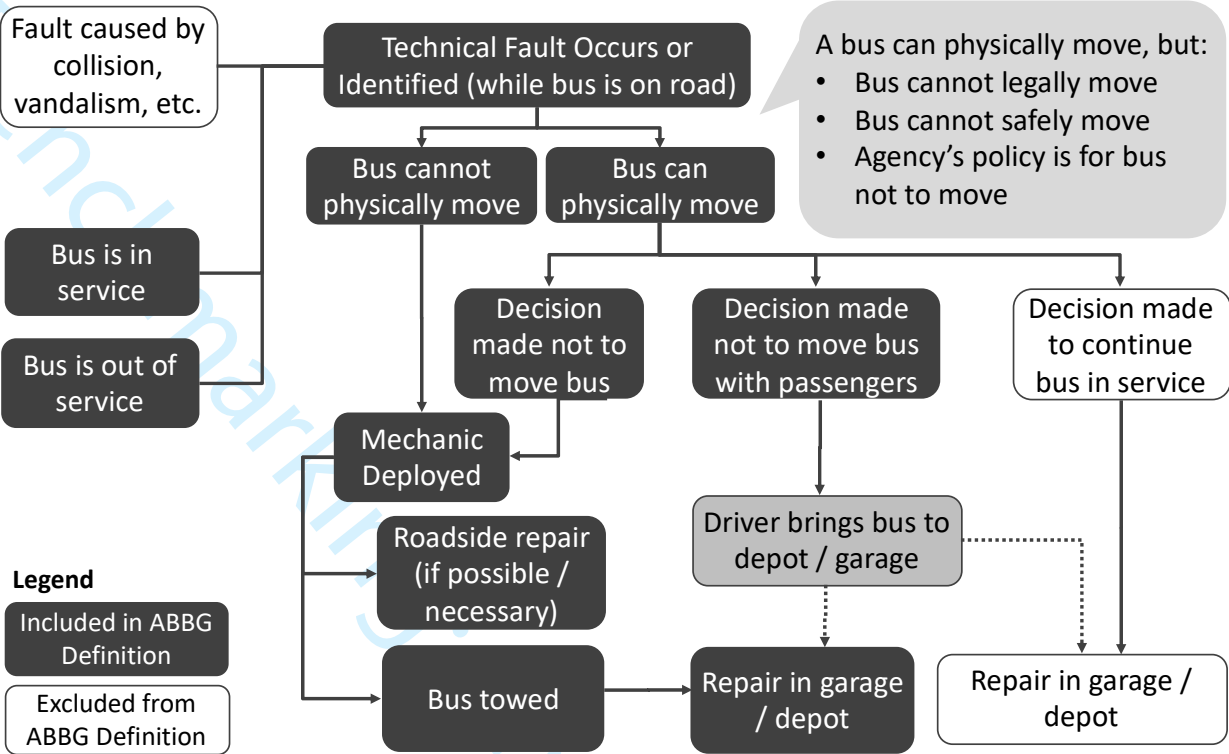


FIGURE 3 Flow Diagram Depicting ABBG Definition of a Road Call due to Technical Faults

The flow diagram in Figure 3 resulted in the following definition for “Number of Road Calls Due to Technical Faults” as agreed by ABBG agencies: The total number of any technical road calls recorded during the year, including road calls due to technical faults that do not directly impact customer service, such as those made during non-revenue service. This includes situations in which the bus is repaired roadside and those in which the bus is driven or towed back to the depot/garage to be repaired. This should include road calls for tire failures, but should exclude "service" road calls, such as those for passenger incidents, collisions, or criminal incidents. Also include change outs due to technical faults, where passengers are moved to another bus, while the bus with a failure returns to the garage (regardless of impact to the passengers).

A common challenge in dealing with data is the possible change in data sources and technology over time. This requires specificity in definition and understanding of when alternative methodologies are being used, documentation of changes, and a balance between preserving meaningful trends and removing previous year data that is no longer comparable. Passenger miles is one data item that often has significant changes in value due to changes in methodology (samples) and technology (APC, customer surveys), but another example is on-time performance. Table 2 lists the drop in performance that occurred for six ABBG organizations when they went from sampling using manual supervisor assessment of on-time performance to data collection using Automatic Vehicle Location (AVL) technology for the entire fleet. As an example, if Agency C previously had reported 90.5% on-time performance using manual supervisor assessment, its performance would have dropped to 82.4% (-8.1%) after

implementation of AVL. The significant impact of AVL on reported performance highlights the need to agencies to manage stakeholder expectations and understanding in advance.

TABLE 2 Drop in On-Time Performance with Implementation of AVL – Trends data from the American Bus Benchmarking Group

Agency	Drop in on-time annual performance after AVL installed (Given as actual change in reported performance)
A	-22%
B	-12.1%
C	-8.1%
D	-8.1%
E	-7.2%
F	-6.9%

Although resource-intensive, ABBG agencies report that the new data items, detailed definitions, and in-depth review of methodologies have led to valuable identification of internal issues and opportunities as well as an improved level of understanding of performance and ways to make changes, as discussed further below under the Data Understanding section.

Annual Standardized Data Collection Methods (2.4, 3.2)

ABBG agencies provide fixed-route data on an annual basis for approximately 130 clearly defined data items. The data is a mix of performance-related data and background data that provide necessary context and understanding. Up to 11 years of data, from 2006 to 2017, were available at the time of this research. Each year, member agencies submit their latest year data as well as any revisions to previous year data that may be necessary due to new methodologies or data systems, the identification of any errors, or the fact that a data item definition has changed per agreement of the group. Thus, the annual collection of data over time is a key contributor to data quality, in providing opportunities for revisions as well as supporting data validation, as discussed below.

Data items are collected from agencies using standardized data sheets, with separate sections for operational, financial, profile, fleet, and staff data. The agreed-upon definitions are included in the data workbooks and any new data items or revised definitions are highlighted when distributed to the ABBG agencies.

One data collection challenge for ABBG is fiscal year definition. Four different fiscal years are used across the ABBG, with fairly good distribution across the groups. During the initial setup of the ABBG, agencies discussed providing calendar year data but agreed that this would be too onerous and result in inaccuracies. Instead, it was agreed that agencies would provide data for their respective fiscal years, as then the benchmarking data is consistent with data that is internally checked and validated. Thus, for those agencies that contributed data to this research, ABBG data were from one of the following time periods (with agencies using that time period indicated):

- July 1 to June 30 (7 agencies)
- October 1 to September 30 (4 agencies)
- January 1 to December 31 (6 agencies)
- April 1 to March 31 (2 agencies)

In-depth Data Validation and Detailed Follow-Up (3.3, 4.5)

The ABBG data validation process, which is an essential piece of ensuring data quality, consists of the following steps, performed for each agency on an annual basis:

- (1) Identification of:
 - a. Missing data items,
 - b. Major changes from previous year data (both in direction and magnitude),
 - c. Internal inconsistencies within that year’s data (for that agency only), and
 - d. Significant differences in performance (direction or magnitude) compared to other agencies.
- (2) Development and issuance of follow-up questions identifying issues and asking for revisions or explanations (if none were provided originally) and subsequent review of responses and additional follow-up as necessary.
- (3) Documentation of explanations whether service-related or methodological.

The repetition of these steps over time adds to data quality and understanding (as discussed below in terms of trend tracking) and is consistent with the iterative, continuous process that is part of the overall ABBG benchmarking approach.

Challenge: Data Understanding

Data understanding in the ABBG is supported by some of the same practices that improve data quality, namely the small group size (2.2), annual cycle (2.4), and in-depth follow-up process (4.5). In addition, there are six sub-elements of Figure 2 that specifically address the challenge of understanding and are discussed in this section.

Bi/multi-dimensional Normalization (4.1)

As mentioned previously and evident in the ABBG fixed-route KPI system (Table 1), nearly all of the KPIs have more than one normalization factor. Trompet *et al.* (2018) provides evidence that conclusions from relative performance differences observed in a single KPI (e.g. only normalized for scale by a single denominator) can be biased. Due to variability between peers (even when all peers are in the United States) of operating characteristics such as commercial speed, trip length, vehicle capacity, and network efficiency (i.e., the amount of deadheading) it is necessary that performance is at least reviewed from two different dimensions, in order to obtain an improved and more balanced understanding of relative performance. This is especially important for bus organizations that have ‘extreme’ values in any of these four identified variable operating characteristics.

In this section, the methodology presented in Trompet *et al.* (2018) to establish variability in service characteristics and understand its implications for performance normalization is applied to 2017 data from up to 19 ABBG agencies for the four operating characteristics identified above as well as two profile characteristics. The results are shown in Table 3, with the characteristics ranked by variability for each of the two categories. The most variable service characteristic by some margin is network efficiency (CV=0.34), defined by the proportion of total vehicle miles that are non-revenue (deadheading) rather than revenue, followed by passenger trip length (CV=0.21). These high variability values highlight the importance of normalizing KPIs in more than one way; for network efficiency, it is important to consider normalizing both per revenue and per total vehicle mile (or per revenue and total vehicle hour), while for trip length, it is important to consider both passenger miles and passenger boardings as normalization factors.

Even with the lower variability ($CV=0.11$) for average commercial speed, the slowest agency (12.1 mph) covers 38% fewer miles per hour than the fastest (19.6 mph), which has a significant impact on relative performance for either of these agencies when a KPI is only normalized by vehicle miles, thus highlighting the importance of also normalizing by vehicle hours. Note that commercial speed is calculated using revenue vehicle hours that exclude layover time (unlike NTD), allowing for a more accurate understanding of in-motion or road speed.

TABLE 3 Variability in Service and Profile Characteristics – American Bus Benchmarking Group 2017 Data

Characteristic	N	μ	Min	Max	σ	CV
<i>Operating Characteristics</i>						
Network efficiency - % of deadheading miles	18	12.1%	5.3%	20.1%	4.1%	0.34
Average passenger trip length (passenger miles per boarding) – miles	19	4.7	3.0	6.5	1.0	0.21
Average commercial speed – mph	19	15.6	12.0	19.0	1.7	0.11
Weighted average vehicle seating capacity	19	38.0	33.5	42.3	2.6	0.07
<i>Profile Characteristics</i>						
Average fleet age	19	8	5.3	11.1	1.7	0.21
Typical hourly driver wage (5 years of experience)	17	\$24.20	\$18.50	\$31.20	\$3.60	0.15

N = Number of bus organizations in sample

μ = Sample average

Min = Minimum value

Max = Maximum value

σ = Standard deviation

CV = Coefficient of variation (σ / μ).

Multi-KPI Analysis (2.2, 4.2)

The ABBG's balanced scorecard framework (Morse *et al.*, 2017; Kaplan and Norton, 1992) and KPI system (see Table 1) is comprehensive and intentionally is not summarized in a composite indexed score, unlike much of the relevant benchmarking literature and exercises (e.g., Al Mamun and Lownes, 2011). This is because the ABBG defines success as good performance across all areas and prioritizes understanding performance differences across these areas. The

ABBG has found that being very good in one area often means performing less well in other areas.

Figure 4 provides a specific example of how, even within one success dimension, a holistic approach combined with bi-dimensional normalization as described above leads to very important conclusions that otherwise would be missed. Figure 4 ranks agencies by their average cost efficiency (cost per output) performance (assessed as: below, the same as, or above the group's average) for seven operating cost KPIs (the first seven rows), representing total and subcomponent operating costs normalized by either vehicle hour or mile, depending on which one is more appropriate to the cost being measured. Within these cost efficiency KPI rows, one can identify subareas of higher or lower cost for individual agencies. The bottom two rows show cost effectiveness (cost per outcome) performance for total operating cost normalized by boardings and passenger miles. These additional two rows demonstrate that some agencies with high cost efficiency (low cost per output) have low cost effectiveness (high cost per outcome), such as Agency C, while other agencies have low cost efficiency (high cost per output) but high cost effectiveness (low cost per outcome), such as Agency R. In addition, this table provides another example why bi-normalization is important in cases where high variability in operating characteristics exists. Relative cost effectiveness performance for Agency F and Agency Q is very different when normalized per passenger boarding or passenger mile. This is a result of the fact that Agency F has long trip lengths, and Agency Q has short trip lengths. This variability in trip lengths is shown in Table 3.

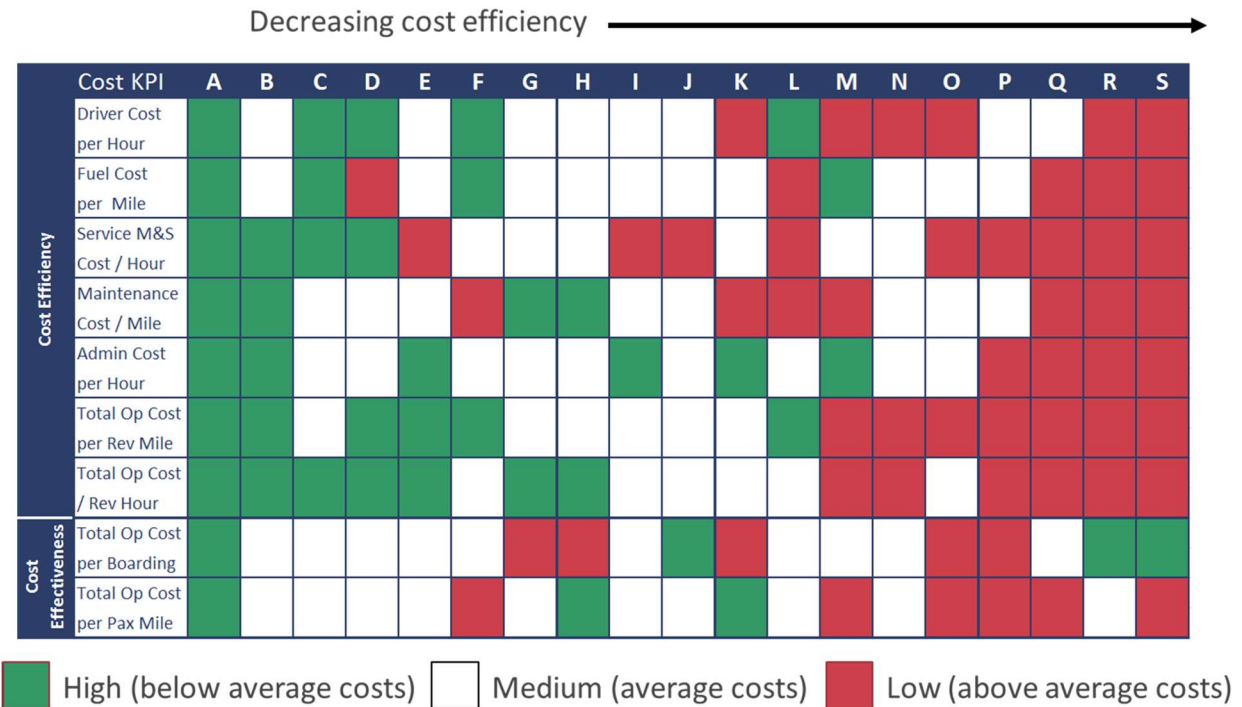


FIGURE 4 A Comparison of Cost Efficiency and Cost Effectiveness Performance – American Bus Benchmarking Group, 2017 Data

Breakdown of Data Items (4.3)

There is a risk when comparing high-level data items of leading to unhelpful or inaccurate generalizations about causation or contributing factors. Of course any further breakdown of data can be resource-intensive so must be worthwhile. The ABBG, as a result of the close working relationship between agencies and the research team, has successfully identified and agreed to a number of worthwhile breakdowns. These include the following:

- (1) 'In-service' vehicle hours broken down into revenue hours and layover hours. This allows for more accurate calculation of revenue speed (excluding layover hours) and also to understand the impact of layover length on driver productivity and on-time performance.
- (2) Fleet utilization broken down into peak fleet utilization, utilization of fleet not in use (maintenance or spare), and fleet utilization by hour of the day (24 hours). The latter in particular allows for an understanding of vehicle and staff utilization, level of service (frequency and coverage), deadheading, and availability of vehicles for maintenance.
- (3) Absenteeism broken down into paid unplanned (primarily sick leave), paid planned (primarily vacation), and unpaid. Paid unplanned leave is the most disruptive to operations but also the most controllable. Paid planned leave depends on state regulations, agency policy, and tenure of staff.
- (4) Vehicle collisions broken down into the categories of preventable, unpreventable, and on-property. Agencies have found the on-property category in particular very helpful to track and benchmark to confirm or identify issues with facilities and training.

Two example KPIs have been included in this paper to demonstrate how the additional breakdown of data can lead to improved understanding of the data. Figure 5 below shows the breakdown of vehicle collision data, with preventable being of most concern, followed by on-property (which are often preventable or at least within the agency's control). Although it is also important to consider collision rates, this graph helps agencies identify where they may have an unusually high proportion of preventable collisions regardless of performance. Thus, even if agency R or S had the lowest collision rates in the group, they may have the greatest potential to reduce their rate further, given the high proportion of preventable collisions (and in the case of agency R, high proportion of on-property collisions). Figure 6 (in the Trend Tracking and Analysis section below) shows the breakdown of absenteeism data for all staff.

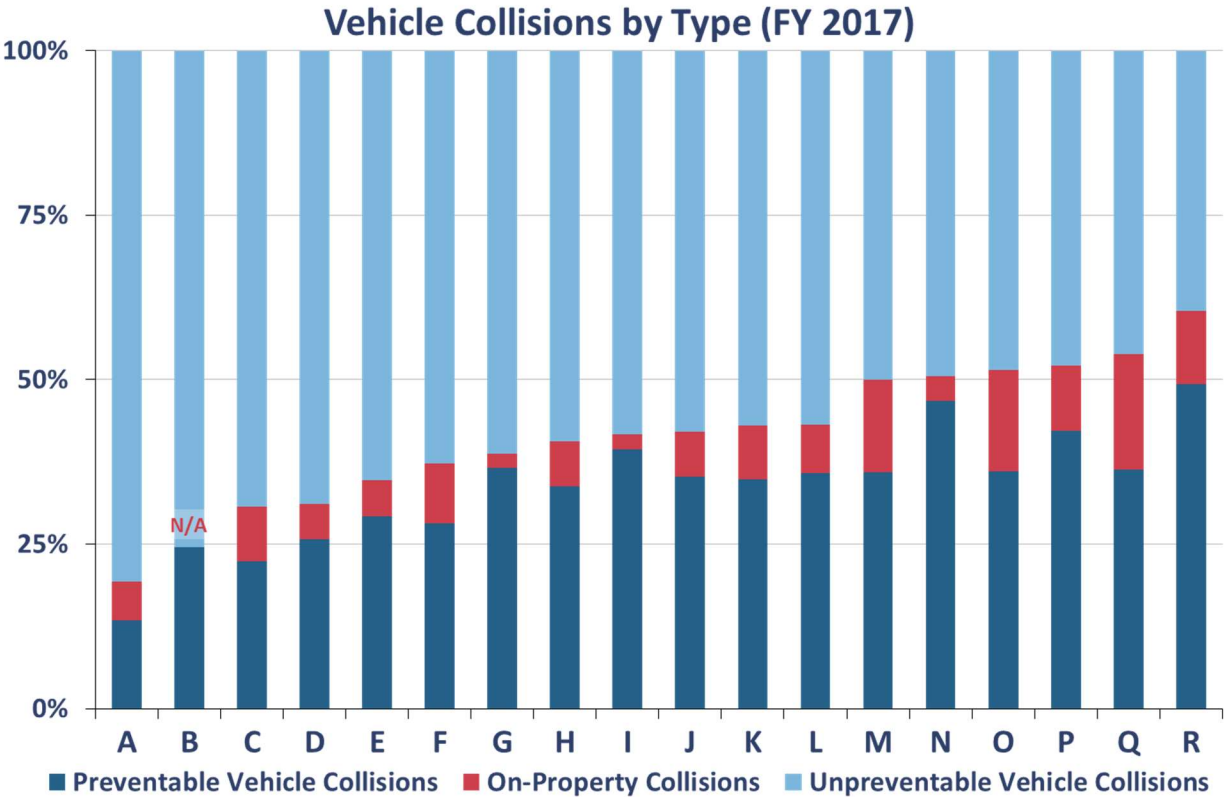


FIGURE 5 Proportion of Collisions due to Preventable, Unpreventable, and On-Property Collisions for the American Bus Benchmarking Group (2017)

Trend Tracking and Analysis (4.4)

Tracking trends across time facilitates data validation as well as the identification of improving or worsening performance. In many cases, other agencies can learn from an agency that is improving in performance regardless of the agency’s relative performance. Trends also inform the agency whether they are going in the right direction or not so they can assess the need for intervention or effectiveness of a change. Trends analysis can also be combined with the breakdown of data into subcomponents. Figure 6 shows one example of this type of analysis from the ABBG, looking at trends in absenteeism for all staff, broken down into the three types as described previously. The ABBG breaks down these data further into staff category (e.g., drivers and maintenance).

In Figure 6, Agency L is fourth highest in total absenteeism but sixth lowest in paid unplanned absenteeism, which has the most direct (and negative) impact on operations. Agency O has faced challenges in paid unplanned absenteeism but has managed to decrease the rate over time. By understanding the information shown in the absenteeism KPI, including the additional information provided by the breakdown of data and tracking trends, ABBG agencies have been able to identify their strengths and weaknesses. They have also been able to share a number of strategies over the years to reduce paid unplanned absenteeism, including providing wellness and fitness facilities and incentives, incorporating absenteeism into performance and promotional reviews, and partnering with unions.

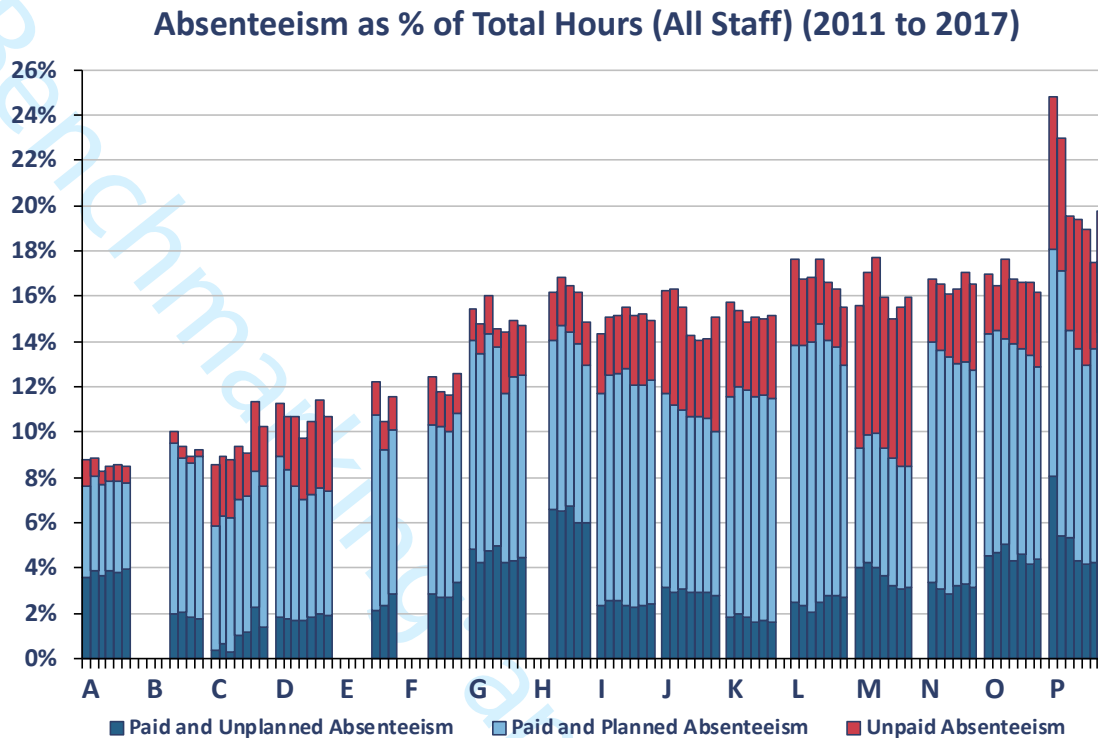


FIGURE 6 Staff Absenteeism Over Time, Broken into Type of Absenteeism for the American Bus Benchmarking Group

Context (2.6)

Understanding relative performance relies on the ability to assess performance within the local and agency context. For the ABBG, differences in state regulatory, economic, and political environments are particularly important to consider given that the ABBG has 24 agencies in 18 states as shown in Figure 1. The ABBG primarily addresses contextual differences through collecting background data, such as the driver wages included in Table 3, and identifying and documenting such differences through agency feedback, the research team's investigations, and small, agency-led topical studies. For example, an ABBG study on absenteeism identified major differences in state Family Medical Leave regulations and agency benefits that were a contributing factor to absenteeism rates (Figure 6).

CONCLUSION

This paper has presented in detail how the ABBG has applied a comprehensive, iterative approach to address the general benchmarking challenges of comparability, data quality, and understanding. Although the application of this research was to mid-sized transit agencies in a U.S. context, it is a model that has been used in other public transport modes and in international groups. Practitioners in public transport and other industries can use this paper as a guide to improve their approaches to peer selection, KPI identification, data collection and analysis, and understanding performance. By covering all the elements presented in the methodology (Figure 2), the ABBG has achieved high-quality, comparable data, leading to improved understanding. As a result, ABBG agencies are able to learn from relative performance, best practices, and lessons learned across the group and apply that learning to their own operations.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

There are several limitations to this research and the framework provided. Benchmarking can be a valuable exercise for agencies in most sectors – but to do it well requires time and commitment. This requirement means that the benchmarking approach described in this paper, as well as any similar research, is most suitably applied to a concise group of agencies (rather than industry-wide). In addition, the data used for this approach are annual and system-wide rather than specific to routes or peak times and thus result in high-level indicators. These indicators consequently require further drill-down analysis by the individual agencies but create a useful starting point. Thus it is strongly advised that participants continuously review the benchmarking results to identify internal actions they can take to improve their performance by applying lessons learned from peers that realized superior performance. Based on ABBG experience, these improvements can take the form of changes in policy, planning, operations, stakeholder communication, and/or program development.

A key research finding and practical implication of this research is that this methodology should not be applied for only for one year, but instead requires a minimum of three years to gain maturity. To ensure relevance over this time and beyond, the KPI system itself (including the definitions) must be continuously reviewed against changing technology, data sources, and priorities, while protecting the comparability of trends where possible. The challenges that this paper focused on (comparability, quality, and understanding) are broad and cross-cutting for all transport and industries. However, it is also important when applying this methodology elsewhere to account for any challenges specific to the chosen participants being benchmarked. In the case of the ABBG, the challenges of the US context needed to be considered, such as different fiscal years, different local and state regulations, and the national requirement for paratransit provision. There is an opportunity for future research to apply this methodology to other industries and in other countries.

The benchmarking methodology used by the ABBG has been designed to be dynamic and continuous. In the near future, this adaptability will be tested both by the opportunity for new metrics that better capture customer impact and by the rise of on-demand services and autonomous vehicles. Future research will be needed to identify how this approach can accommodate these changes and which adjustments will be needed.

REFERENCES

- Al Mamun, M. S., and Lownes, N. E. (2011), "A Composite Index of Public Transit Accessibility", *Journal of Public Transportation*, Vol. 14 No. 2, pp. 69-87. Available at: <http://doi.org/10.5038/2375-0901.14.2.4>.
- Alstete, J. (2008), "Measurement benchmarks or "real" benchmarking? An examination of current perspectives", *Benchmarking: An International Journal*, Vol. 15 No. 2, pp. 178-186. Available at: <https://doi.org/10.1108/14635770810864884>.
- Chu, X., Fielding, G. J., and Lamar, B. W. (1992), "Measuring Transit Performance Using Data Envelopment Analysis", *Transportation Research Part A: Policy and Practice*, pp. 223-230. Available at: [https://doi.org/10.1016/0965-8564\(92\)90033-4](https://doi.org/10.1016/0965-8564(92)90033-4).
- Cook, T. J., Collins, D. G., and Monast, K. (2010), *Benchmarking Guidebook for North Carolina Community Transportation Systems*. North Carolina Department of Transportation. Available at: <https://itre.ncsu.edu/wp-content/uploads/2016/03/FINAL-Short-CT-Guidebook-20110222.pdf>.
- Fielding, G., Babitsky, T., and Brenner, M. (1985), "Performance evaluation for Bus Transit", *Transportation Research Part A, Vol. 19A* No. 2, pp. 73-82. Available at: [https://doi.org/10.1016/0191-2607\(85\)90009-3](https://doi.org/10.1016/0191-2607(85)90009-3).
- Florida Department of Transportation (2014), *Best Practices in Evaluating Transit Performance: Final Report*. Tallahassee, FL: Freight Logistics and Passenger Operations, Transit Office. Available at: https://fdotwww.blob.core.windows.net/sitefinity/docs/default-source/content/transit/pages/bestpracticesinevaluatingtransitperformancefinalreport.pdf?sfvrsn=48878730_0.
- Fu, L., and Xin, Y. (2007), "A New Performance Index for Evaluating Transit Quality of Service", *Journal of Public Transportation*, Vol. 10 No. 3, pp. 47-69. Available at: <http://doi.org/10.5038/2375-0901.10.3.4>.
- Gan, A. F., Gui, F., and Tang, L. (2011) 'System for Transit Performance Analysis Using the National Transit Database', *Journal of Public Transportation*, Vol. 14 No. 3, pp. 87-107. Available at: <http://doi.org/10.5038/2375-0901.14.3.5>.
- Geerlings, H., Klementschtz, R., and Mulley, C. (2006) 'Development of a methodology for benchmarking public transportation organisations: a practical tool based on an industry sound methodology', *Journal of Cleaner Production*, 14, pp.113-123. Available at: <https://doi.org/10.1016/j.jclepro.2005.03.021>.
- Hartgen, D. T., and Horner, M. W. (1997), *Report 163: Comparative Performance of Major U.S. Bus Transit Systems: 1988—1995*. Charlotte, NC. University of North Carolina.
- Hendren, P. (2011), "Bringing Performance-Based Management to Transit". In *Performance Measure of Transportation Systems: Summary of the Fourth International Conference* (K. Turnbull, ed.) (pp. 66-69). Washington, DC: The National Academies Press. Available at: <https://doi.org/10.17226/22535>.
- Henning, T., Essakali, D., and Oeh, J. E. (2011), *A framework for urban transport benchmarking*. Washington DC: World Bank. Available at: <http://www.utbenchmark.in/img/RefDocuments/Home-Ref-1-3-1.pdf>.

- Hensher, D., and Daniels, R. (1995), "Productivity Measurement in the Urban Bus Sector", *Transport Policy*, Vol. 2, pp. 179-195. Available at: [https://doi.org/10.1016/0967-070X\(95\)96748-A](https://doi.org/10.1016/0967-070X(95)96748-A)
- Hilmola, O. (2011), "Benchmarking efficiency of public passenger transport in larger cities", *Benchmarking: An International Journal*, Vol. 18 No. 1, pp. 23-41. Available at: <https://doi.org/10.1108/14635771111109805>.
- Hinton, M., Francis, G. and Holloway, J. (2000), "Best practice benchmarking in the UK", *Benchmarking: An International Journal*, Vol. 7 No. 1, pp. 52-61. Available at: <https://doi.org/10.1108/14635770010314954>.
- Kaplan, R. S., and Norton, D. P. (1992), "The Balanced Scorecard – Measures that Drive Performance", *Harvard Business Review*, 7 (1), pp. 71-79.
- Karlaftis, M. G. (2004), "DEA approach for evaluating the efficiency and effectiveness of urban transit systems", *European Journal of Operational Research*, Vol. 152 No. 2, pp. 354-364. Available at: [https://doi.org/10.1016/S0377-2217\(03\)00029-8](https://doi.org/10.1016/S0377-2217(03)00029-8).
- Kassens-Noor, E., Kotval-Karamchandani, Z., Brush, A., Doshier, K., and Biskey, M. (2019), "Michigan's public transportation: An application of statewide performance assessment and management", *Transportation Research Interdisciplinary Perspectives*, 1. Available at: <https://doi.org/10.1016/j.trip.2019.100013>
- Kittelsohn & Associates, Inc., Urbitrans Inc., LKC Consulting Services, Inc., MORPACE International Inc., Queensland University of Technology, Y.A. Nakanishi (2003), *The Transit Cooperative Research Program (TCRP). Report 88: A Guidebook for Developing a Transit Performance-Measurement System*. Washington, D.C.: Transportation Research Board. Available at: http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_rpt_88.pdf.
- Lema, N., and Price, A. (1995), "Benchmarking: performance improvement towards competitive advantage", *Journal of Management in Engineering*, 11(1), pp 28-37.
- Magd, H. and Curry, A. (2003), "Benchmarking: achieving best value in public-sector organisations", *Benchmarking: An International Journal*, Vol. 10 No. 3, pp. 261-286. Available at: <https://doi.org/10.1108/14635770310477780>.
- Meyer, M. D. (2002), "Measuring System Performance: Key to Establishing Operations as a Core Agency Mission", *Transportation Research Record*, Vol. 1817 No. 1, pp. 155-162. Available at: <https://doi.org/10.3141/1817-20>
- Min, H., and Ahn, Y. (2017), "Dynamic Benchmarking of Mass Transit Systems in the United States Using Data Envelopment Analysis and the Malmquist Productivity Index", *Journal of Business Logistics*, Vol. 38 No. 1, pp. 55-73. Available at: <https://doi.org/10.1111/jbl.12148>.
- Morse, L., Trompet, M., Barron, A., and Graham, D. (2017), "Development of a Key Performance Indicator System to Benchmark Relative Paratransit Performance", *Transportation Research Record*, pp. 1-8. Available at: <https://doi.org/10.3141/2650-01>.
- Perk, V., and Kamp, N. (2004), *Benchmark Rankings for Transit Systems in the United States*. Tallahassee, FL: State of Florida Department of Transportation. doi:NCTR-527-03/FDOT BC137-43. Available at: <https://www.nctr.usf.edu/pdf/527-03.pdf>.

Ryus, P. et al. (2010), *TCRP Report 141: A Methodology for Performance Measurement and Peer Comparison in the Public Transportation Industry*. Washington, D.C.: Transportation Research Board. Available at: <https://doi.org/10.17226/14402>.

Trompet, M., Anderson, R. J., and Graham, D. J. (2009), "Variability in Comparable Performance of Urban Bus Operations", *Transportation Research Record: Journal of the Transportation Research Board*, pp. 177-184. Available at: <https://doi.org/10.3141/2111-20>.

Trompet, M., Anderson, R. J., and Graham, D. J. (2018), "Improved understanding of the relative quality of bus public transit using a balanced approach to performance data normalization", *Transportation Research Part A: Policy and Practice*, Vol. 114, pp. 13-23. Available at: <https://doi.org/10.1016/j.tran.2017.11.019>.

Trompet, M., Parasam, R., and Anderson, R. J. (2013), "Benchmarking Disaggregate Customer Satisfaction Scores of Bus Operators in Different Cities and Countries", *Transportation Research Record*, pp. 14-22. Available at: <https://doi.org/10.3141/2351-02>.

Verbich, D., Badami, M., and El-Geneidy, A. (2017), "Bang for the buck: Toward a rapid assessment of urban public transit from multiple perspectives in North America", *Transport Policy*, Vol. 55, pp. 51-61. Available at: <https://doi.org/10.1016/j.tranpol.2016.12.002>.

Zairi, M. (1994), "Benchmarking: The Best Tool for Measuring Competitiveness", *Benchmarking for Quality Management & Technology*, Vol. 1 No. 1, pp. 11-24. Available at: <https://doi.org/10.1108/14635779410056859>.

A BENCHMARKING FRAMEWORK FOR UNDERSTANDING BUS PERFORMANCE IN THE U.S.

ABSTRACT

Purpose This paper describes a benchmarking framework applied to medium-sized urban public bus agencies in the United States which has overcome the challenges of data quality, comparability and understanding.

Design/methodology/approach The benchmarking methodology described in this paper is based on lessons learned through seven years of development of a fixed route key performance indicator (KPI) system for the American Bus Benchmarking Group (ABBG). Founded in 2011, the ABBG is a group of public medium-sized urban bus agencies that compare performance and share best practices with peers throughout the United States. The methodology is adapted from the process used within international benchmarking groups facilitated by Imperial College and consists of four main elements: peer selection, KPI system development, processes to achieve high-quality data, and processes to understand relative performance and change.

Findings The four main elements of the ABBG benchmarking methodology consist of eighteen sub-elements, which when applied overcome three main benchmarking challenges; comparability, data quality, and understanding. While serving as examples for the methodology elements, the paper provides specific insights into service characteristics and performance among ABBG agencies.

Research limitations/implications The benchmarking approach described in this paper requires time and commitment and thus is most suitably applied to a concise group of agencies.

Practical implications This methodology provides transit agencies, authorities and benchmarking practitioners a framework for effective benchmarking. It will lead to high-quality comparable data and a strong understanding of the performance context to serve as a basis for organizational changes, whether for policy, planning, operations, stakeholder communication, or program development.

Originality/value The methodology, while consistent with recommendations from literature, is unique in its scale, in-depth validation and analysis, and holistic and multi-dimensional approach.

Keywords: performance, performance management, performance measurement, benchmarking, quality, balanced scorecard, continuous improvement

Research paper

ACKNOWLEDGEMENTS

The authors thank the member agencies of the American Bus Benchmarking Group for their willingness to share their experiences with the development of fixed route benchmarking with the wider transport and benchmarking industry.

INTRODUCTION

Benchmarking performance can be a powerful tool in management. It can provide identification and understanding of strengths and weaknesses, improve processes, and lead to establishing priorities and realistic targets. However, for such benchmarking to be effective, it must be based on high-quality comparable data and a strong understanding of the performance context. The joint development of a fixed-route bus key performance indicator (KPI) system by the members of the American Bus Benchmarking Group (ABBG) and Imperial College London offers an effective methodology for benchmarking bus transit performance. The ABBG's founding agencies wanted to go beyond the scope of the Federal Transit Administration's National Transit Database and adopt a benchmarking methodology used successfully internationally.

Founded in 2011, the ABBG is a group of public medium-sized urban bus agencies that agreed to compare performance and share best practices with peers in similar cities throughout the United States. The ABBG is modeled on the International Bus Benchmarking Group (IBBG), a global group of 16 urban bus systems that was founded in 2004 (Trompet *et al.*, 2009; Trompet *et al.*, 2018).

As of 2020, the ABBG is comprised of 24 agencies, but at the time of this research in 2018, data was available from 19 member agencies, as shown in Figure 1. Performance data from six agencies (indicated with a * and italics) were not yet available at the time of writing, as these organizations joined the ABBG within the past two years. Data was available from WeGo Nashville (indicated with a ^), that was an ABBG member at the time but has since left.

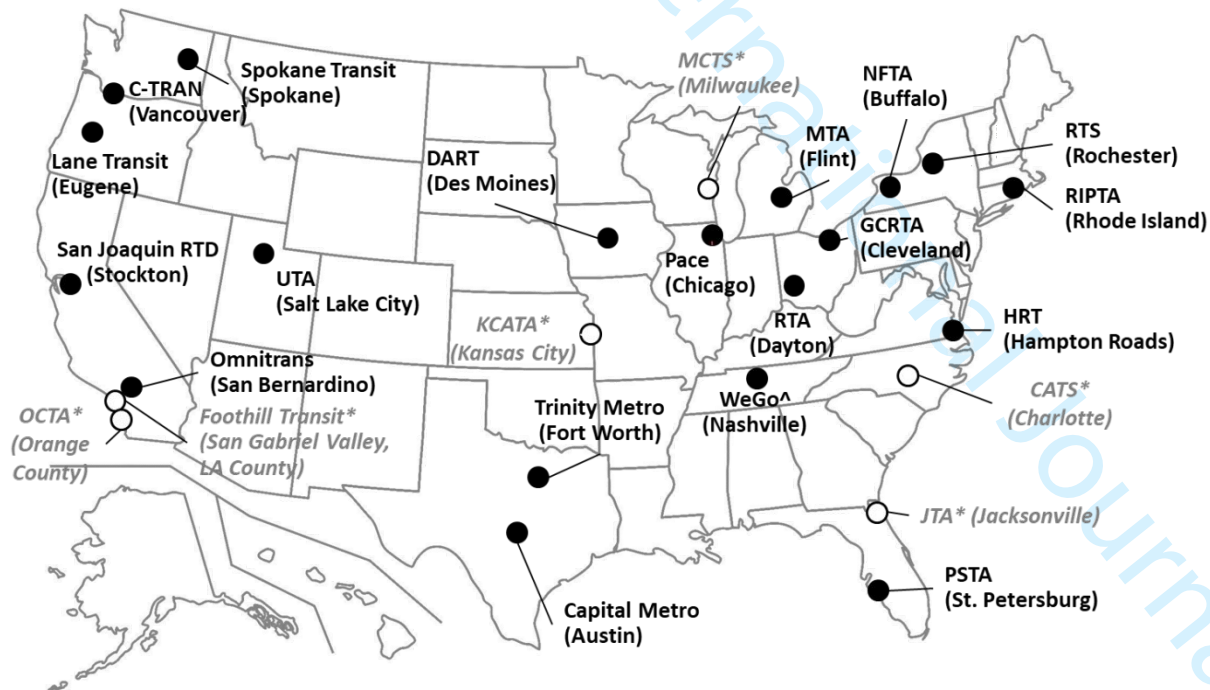


FIGURE 1 Map of ABBG Members (2020) and Research Participants (2018)

Building upon years of international benchmarking experience (especially from the IBBG) and input from ABBG agencies, a fixed-route KPI system was developed from 2011-2014, and dynamically adjusted in the following years to remain in line with agency priorities and technology developments. During the course of development, the following research questions were addressed: what challenges need to be overcome to achieve high-quality comparable data and a strong understanding of the performance context? How can these challenges be addressed in a benchmarking methodology? How can these challenges be addressed within a U.S. context?

Three general categories of challenges were identified that needed to be overcome for the benchmarking system to work well and deliver the desired outcomes (namely, the ability to learn and implement change): comparability, data quality, understanding.

The comparability of data starts with the comparability of the source – in this case, transit agencies – and carries through to comparability of the KPIs selected. Transit agencies can vary in scale, level of service throughout the day and week, management, and types of services. The ABBG’s eligibility criteria and clearly defined scope of benchmarking limit variability and ensure a useful level of similarity while also allowing for benefits from diversity (Trompet *et. al*, 2009). The ABBG’s agency-led prioritization of data and selection of KPIs ensure the benchmarking stays in line with strategic goals.

Achieving high-quality data requires agency commitment, robust definitions and standardized data collection, and in-depth validation processes to ensure accuracy and address differences in data collection and management, which change over time with technology and resources.

Improved data understanding leads to improved understanding of performance. The ABBG approach supports understanding by breaking down data items into subcomponents, using bi/multi-dimensional normalization and multi-KPI analysis, and considering trends and contextual information about the transit agencies’ operating environments and service characteristics.

For the purposes of the U.S. context, specific challenges around different state (and local) contexts, provision of separate paratransit services, and variety of fiscal years also had to be addressed.

For each of the three areas above, this paper describes the challenges and respective solutions identified during the development of the ABBG fixed-route KPI system. To facilitate the discussion, the remainder of the paper is organized as follows: The literature review provides an overview of available relevant academic research and government publications with a focus on U.S. transit performance benchmarking. The methodology section that follows describes the ABBG performance benchmarking process and identifies how elements of the methodology relate to one or more of the three identified challenge areas. The discussion section of the paper then details how these process elements and their respective sub-elements address the three areas of benchmarking challenges listed above, using specific performance and operating characteristic examples from the ABBG. Finally, the paper will provide guidance on how to apply this research, discuss its limitations, and identify opportunities for future work and research.

LITERATURE REVIEW

Performance benchmarking has a rich, robust set of literature, applied to all industries and to all modes of transport (e.g., Hinton *et al.*, 2000). It is worth recognizing that how benchmarking is defined can vary significantly (e.g., as discussed in Alstete, 2008), and that many frameworks

have been identified as successful (e.g., Zairi, 1994 and Geerlings *et al.*, 2006). For public transport specifically, the extensive performance metrics provided by several European Union transport benchmarking projects are often used as a starting point (e.g., EQUIP as covered by Geerlings *et al.*, 2006). Other publications have identified multiple categories of performance indicators such as the eight identified by Kittelson and Associates, Inc. *et al.* (2003) including service delivery, community impacts, travel times, safety and security, maintenance and construction, economics, paratransit, and comfort. Kassens-Noor *et al.* (2019) includes a review of the development of technology-driven performance measures in transit. Perk and Kamp (2004) review various previous attempts to identify composite scores or indexes such as the Transit Performance Index developed by the American Public Transportation Association. Transit service quality and accessibility indices are also common (Fu and Xin, 2007; Al Mamun and Lownes, 2011; Hartgen and Horner, 1997).

The wide body of international literature on transit benchmarking relevant to this research is documented by Trompet *et al.* (2009; 2018). These two publications draw upon this literature as well as the benchmarking experience of the IBBG and provide two key lessons that have been applied to ABBG's methodology. Trompet *et al.* (2009) found benchmarking to be useful and justifiable because there is sufficient variability in performance between agencies due to factors within management control rather than only those resulting from differences in external factors, thus allowing for agencies to learn lessons from one another. Trompet *et al.* (2018) concluded that due to variability of service characteristics between peers, such as commercial speed or trip length, it is necessary that performance be at least reviewed from two different dimensions in order to understand the normalization bias resulting from this variability and to obtain an improved and more balanced understanding of relative performance. These findings are key to the success of the ABBG as described further in this paper.

In Europe, European Union projects such as EQUIP provided both performance metrics and step-by-step approaches (e.g., Geerlings *et al.*, 2006) to follow. In the U.S., there is a well-established body of government-sponsored publications providing both guidance and documentation of fixed route transit performance-measurement systems and peer comparison (Kittelson and Associates, Inc. *et al.*, 2003; Ryus *et al.*, 2010; Florida Department of Transportation, 2014; Perk and Kamp, 2004; Cook *et al.*, 2010). Although the benchmarking methodology described in this paper is consistent with many of the recommendations in these publications, it is distinct in its scale, in-depth validation and analysis, and holistic and multi-dimensional approach. For example, the step-by-step process used in this research shares the cyclical nature and emphasis on learning (and acting upon that learning) of EQUIP as described by Geerlings *et al.* (2006) but is quite different in its starting point. EQUIP begins with a self-assessment rather than a collaborative process to develop common measures. One of the most relevant U.S. benchmarking projects is a recent statewide effort by the Michigan Department of Transportation to measure performance across 80 rural and urban transit systems (Kassens-Noor *et al.*, 2019). Although its goals and performance metrics are quite different, the project also found that its approach also required significant resources and time, trust between participating entities, and one-on-one individual follow-up.

The Federal Transit Administration's National Transit Database (NTD) serves as the main data source for nearly all US public transit benchmarking literature, especially since it was compiled online in an interactive format that allows for peer selection (Gan *et al.*, 2011). The NTD is a valuable large-scale depository of data for the entire U.S. transit industry that began in the 1970s. Submitting data to the NTD is mandatory for nearly all US transit agencies that

receive federal funding and is the basis for formula allocations of federal transit funds. Although fairly comprehensive, the NTD faces challenges in data availability (Verbich *et al.*, 2017) and data quality and comparability, mainly due to its scale, lack of verification, and lack of context and interpretation (Hendren, 2011).

Drawing upon NTD data, many academic publications covering U.S. transit benchmarking focus on algorithms to improve operational efficiency in terms of resource utilization (vehicles and labor) or high-level evaluations of financial efficiency across a large set of agencies (Fielding *et al.*, 1985). One of the well-established quantitative approaches, using NTD data, is data envelopment analysis (DEA). DEA focuses on measuring and comparing efficiency and effectiveness for a very limited number of inputs (e.g., cost, labor, and vehicles) compared to outputs (e.g., boardings, vehicle miles) (e.g., Chu *et al.*, 1992; Karlaftis, 2004). Min and Ahn (2017) applied DEA to over 200 agencies in the US over a multi-year period. Although commendable for adding trends analysis and consideration of important explanatory context (namely organization type, modes, service area density, and residing state), their analysis is limited to one financial and one service efficiency measure.

When benchmarking public transport, the audience is a key consideration. Much of the existing public transport benchmarking literature assumes the policymaker, taxpayer, or transit user (customer) perspective, although Verbich *et al.* (2017) tries to cover all perspectives including the transit agency, Kassens-Noor *et al.* (2019) considers both the transit agencies and the state government, and Hilmola (2011) applies DEA to public transit benchmarking from the city’s perspective. Policymakers are interested in outcomes such as accessibility and affordability (Henning *et al.*, 2011) or environmental emissions and modal share (Hilmola, 2011); taxpayers in cost efficiency and effectiveness (Min and Ahn, 2017); customers in journey time, comfort, and safety (Verbich *et al.*, 2017). The benchmarking methodology described in this paper is primarily intended for the transit agency but attempts to cover performance areas that are priorities for all stakeholders.

METHODOLOGY

Based on a definition by Lema and Price (1995) benchmarking is defined by the ABBG and its sister benchmarking groups as *a systematic process of continuously measuring, comparing and understanding organizations’ performance and change in performance of a diversity of key business processes against comparable peers to gain information which will help the participating organizations to take action to improve their performance*. These principles are repeatedly identified as success factors in benchmarking literature, such as Magd and Curry (2003).

The development of ABBG’s fixed-route benchmarking process and KPI system adopted the methodology taken for the other benchmarking groups facilitated by Imperial College London. Establishing a robust KPI system that uses a comparable, high-quality dataset and obtaining the understanding of underlying differences in performance takes time. While benefiting from lessons learned and experience from the IBBG, it took the ABBG three years of iterative definition development, data collection and analysis before the member agencies were sufficiently satisfied with the level of comparability to be able to use the data for performance comparison. Trompet *et al.* (2009) summarized the necessary benchmarking conditions and decisions necessary in those development years to gradually increase the quality of the dataset.

The methodology that has been the foundation for the establishment, development and continuation of successful benchmarking groups such as the IBBG and applied to the ABBG is presented in Figure 2.

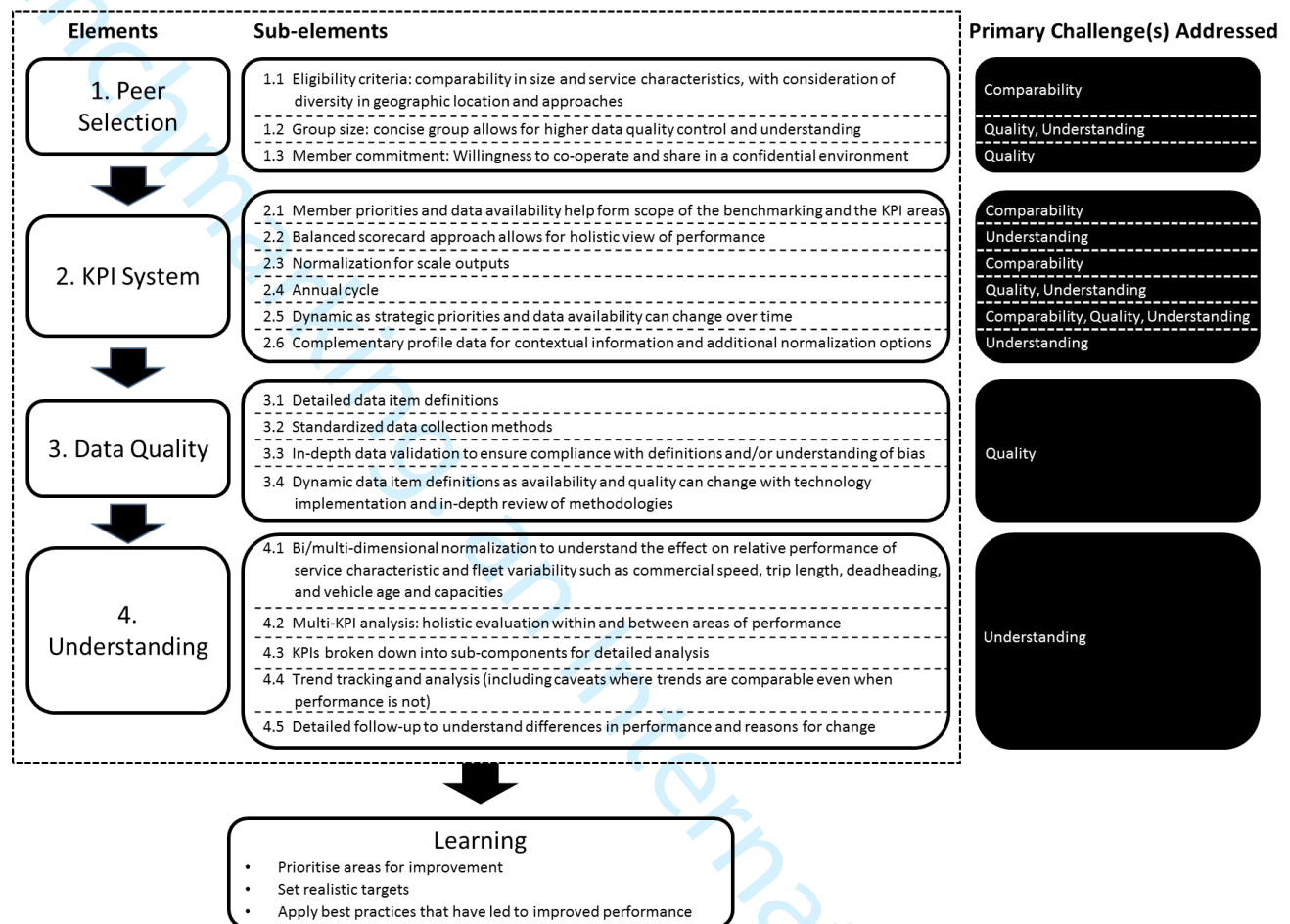


FIGURE 2 ABBG Benchmarking Methodology with Identification of Challenges Addressed by Sub-Element

This methodology consists of four main elements and eighteen sub-elements which, when executed well, enables organizations to learn from the benchmarking and to take action to improve their performance, as expressed by the 'learning' box in the figure. The main elements of the methodology are: peer selection, development of the KPI system, processes to achieve high-quality data, and the analyses and processes to understand relative performance and reasons for change in performance. At the inception of a benchmarking group, these main elements occur in order as expressed through the arrows; however, once the benchmarking group is in the later stages of development and/or established, these main elements occur simultaneously to continuously improve the usefulness and quality of the benchmarking process.

The objective of this paper is to describe how the main elements and respective sub-elements of this benchmarking methodology address the three identified challenges of comparability, data quality, and understanding in benchmarking. Figure 2 outlines how the eighteen sub-elements are linked to one, two, or all three of the benchmarking challenges. The discussion section of this

paper describes these sub-elements and details how each of these methodology elements and sub-elements contribute to addressing the three challenges that needed to be overcome for the benchmarking system to work well and deliver the desired outcomes.

DISCUSSION OF RESULTS

This section is presented in three subsections, each focused on one of the challenges from Figure 2 and the sub-elements (with numbered references given in parentheses) that address the respective challenge.

Challenge: Comparability

This section describes how four of the sub-elements from Figure 2 address comparability.

Eligibility Criteria (1.1)

The ABBG membership criteria, developed by the ABBG founding members, are as follows:

- (1) Mid-sized US or Canadian bus organization,
- (2) Between approximately 5 million and 50 million annual boardings,
- (3) Between approximately 100 and 600 vehicles in fleet,
- (4) Similar service and operating characteristics (fixed-route urban or suburban ‘local’ bus service with all-day, every-day operation with at least 15-30 minute headways on key routes),
- (5) Majority public operation preferred and public management / administration required,
- (6) Willingness to invest resources and fully participate, and
- (7) Geographic diversity.

Generally, the ABBG’s eligibility criteria are designed to balance similar service characteristics (criterion 4) while allowing for variability (criteria 2, 3, and 7) that enables diversity and exchange of new ideas, as documented in Trompet *et al.* (2009). Although there are benefits from benchmarking across multiple countries, the focus on North America (criterion 1) means greater consistency in terminologies, data, operating and regulatory environments, and in the scope of services, as covered below. Differences in scale can be addressed through normalization factors, so the limitations in size (criteria 2 and 3) are due to contextual differences between larger and smaller cities and practical considerations as described under Data Quality. Criteria 5 and 6 are intended to ensure the agency is both able and willing to provide quality data; given that the ABBG was founded to go beyond the NTD, agencies need to have the resources and ability (in terms of technology and data systems) to provide both more and different data than for NTD to meet the definitions agreed upon by the group. The importance of this type of commitment has been documented in benchmarking literature (e.g., Magd and Curry, 2011).

The criteria are applied only to interested parties and then membership is subject to the group’s consensus; the group also has a limitation on size as discussed in the next section.

Scope of Benchmarking (2.1)

To ensure comparability of performance the scope of the ABBG benchmarking effort had to be defined. The group agreed that all regularly scheduled services open to the general public were included in the scope of the ABBG fixed route benchmarking. These include the NTD categories of Motor Bus (MB), Commuter Bus (CB), Trolley Bus (TB), and Bus Rapid Transit (RB), as well as circulators and routes that serve schools or universities but are open to the general public.

In addition, the group agreed consistent with NTD to include route-deviation/flexible services, including those that serve zones, in the fixed route data set. These services are common among the ABBG agencies and considered 'core business'. Excluded services are any special event services or rail replacement.

Unlike in most of the world, all public transit agencies in the U.S. are inherently multimodal because of the requirement for complementary paratransit service, which has very different service characteristics (and thus benchmarking needs) than fixed route service. The ABBG has developed a distinct, separate paratransit benchmarking process and KPI system as described in Morse *et al.* (2017). This separate system complements the fixed route system in terms of both understanding (see below) and comparability, because it accounts for all shared costs and resources across the two data sets and ensures they are allocated between the two modes, especially for those agencies that share drivers and/or maintenance. In addition, seven of the ABBG agencies operate one or more other mode (e.g., light rail, commuter rail, etc.). This has required careful allocation of shared resources (e.g., labor), costs, and revenue.

KPI System (2.1, 2.3, 2.5)

The foundation of the KPI system is a balanced scorecard across four dimensions: Growth & Learning, Internal Processes, Customer, and Financial (Kaplan and Norton, 1992), adapted to transit with the addition of Safety and Environment dimensions as described in Morse *et al.* (2017). Within each success dimension, KPIs were then developed based on agency priorities and data availability or willingness to collect the required data. The fact that the KPIs are agreed by agency senior executives adds to improving the comparability of the KPI system, as KPIs are selected which reflect common interests, challenges, and requirements. Furthermore, the KPI system is dynamic; therefore, the KPI system can be adjusted to reflect changes in strategic goals and/or data availability. For example, it is expected that KPIs that focus on actual customer impact (such as customer delay minutes) will be adopted by the ABBG once technology allows for their measurement. The ABBG fixed-route KPI system at the time of this research (2018) is shown in Table 1. The ABBG also conducts a separate, complementary Customer Satisfaction Survey benchmarking effort (Trompet *et al.*, 2013) as recommended by several studies (e.g., Meyer, 2002).

TABLE 1 ABBG Fixed Route Key Performance Indicator System

Success Dimensions	Key Performance Indicators
Growth & Learning	Ridership: Passenger Boardings (<i>5-year % change</i>) Service Levels: Vehicle Revenue Miles and Hours (<i>5-year % change</i>) Passengers per Revenue Mile & Hour Staff Training
Customer	Customer Information (scheduled and real-time) On-Time Departure Performance Passenger Miles per Revenue Capacity Mile Passenger Miles per Revenue Seat Mile Lost Vehicle Mile Missed Trips
Internal Processes	Peak Fleet Utilization Network Efficiency (revenue miles & hours per total miles & hours, non-revenue split by category) Staff Productivity (total vehicle hours & miles per labor hour, overall and by category) Staff Absenteeism Rate (by staff category and absenteeism type) Fleet Reliability (<i>miles/time between road calls due to technical faults</i>)
Financial	Total Cost per Vehicle Mile & Hour Total Operating Cost per Vehicle Mile and Hour (<i>service operation, maintenance, administration</i>) Service Operation Cost per Revenue Mile & Hour Total Operating Cost per Boarding & Passenger Mile Operating Cost Recovery Fare Revenue per Boarding & Passenger Mile
Safety	Number of Vehicle Collisions per Vehicle Mile and Hour (<i>preventable, non-preventable, on-property</i>) Number of Staff Injuries per Staff Work Hour Number of Passenger Injuries per Boarding & Passenger Mile Number of 3 rd Party Injuries per Vehicle Mile & Hour
Environmental	Fuel Consumption (<i>per total vehicle mile, passenger mile, & capacity mile</i>) Carbon Dioxide Emissions (<i>per total vehicle mile & passenger mile</i>)

Each of the KPIs listed in Table 1 is normalized for scale outputs, which is arguably the sub-element of the system that is most relevant to comparability. The ABBG uses a variety of scale denominators, but there are four primary ones, two on the demand side (passenger boardings and passenger miles) and two on the supply side (vehicle miles and hours, both total and revenue). While most benchmarking exercises and literature focus on the use of mileage as a scale denominator, useful discussions have been provided on the use of vehicle hours as a preferred denominator over vehicle miles (Fielding *et al.*, 1985; Hensher and Daniels, 1995) as much of cost, revenue, and labor are linked to hours rather than miles. The ABBG therefore uses both supply side output denominators.

In addition, specific KPIs were designed to make data more comparable even when there are significant differences between agencies for both service and fleet characteristics. Two examples of such KPIs are passenger miles per revenue capacity mile, a customer utilization KPI that takes into account both the seating and standing capacity of fleets, and carbon dioxide emissions, an environmental KPI that allows for comparison across fleets even with the diversity of fuels used (diesel, diesel hybrid, gasoline, compressed natural gas, electric trolleybus, electric battery bus).

Challenge: Data Quality

This section describes how eight of the sub-elements from Figure 2 address data quality.

Group Principles (1.2, 1.3)

The ABBG has adopted several core principles, including collaboration, depth, speed, and confidentiality. As documented in Trompet *et al.* (2009), one of the main contributing factors to data quality is an open and honest sharing environment that is supported by a strict confidentiality agreement, which ABBG has adopted. However, the group deliberately made the decision to support this publication and share some of its lessons and information with the broader transit industry. Thus, while the graphs and tables in this paper are still anonymized and/or aggregated, they include actual averages, ranges, and in one case (Figure 6), actual KPI values.

These core principles rely on limiting the size of the group to a maximum number that can comfortably meet in person around one table, to facilitate direct discussions and joint decision-making. The limitation to size also makes in-depth validation more manageable. There is also a minimum benchmarking group size required to ensure sufficient comparability within any one KPI, as documented in Trompet *et al.* (2009). From its extensive experience in benchmarking, the research team has determined the ideal size to be somewhere between 8 and 18 agencies. The ABBG has grown beyond this size due to the group's desire to accommodate additional interest and as a consequence, it is in the process of adopting new protocol, including meeting as two subgroups, to ensure the ABBG principles are still met.

Detailed and Dynamic Data Item Definitions (3.1, 3.4)

The ABBG has an advantage compared to international benchmarking in that nearly all U.S. transit agencies are required to submit comprehensive data to the NTD. However, as mentioned previously, the founding ABBG agencies identified a need to go beyond the NTD in terms of quantity and granularity of data and detail of definitions.

To achieve high quality data, all data items necessary to create the KPIs as listed in Table 1 have been defined carefully. Defining and agreeing definitions is a process that requires research and subsequent discussion amongst member agencies. Figure 3 provides an example of how ABBG approaches defining data items. It is a flow diagram showing how the group has agreed to define technical road calls to measure maintenance reliability. All possible aspects of a road call due to technical faults are listed; however, only those elements in black boxes were selected to be included into the ABBG definition. This diagram was developed based on two surveys of ABBG agencies, detailed follow-up with maintenance staff, and lessons learned from a similar process in the IBBG.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

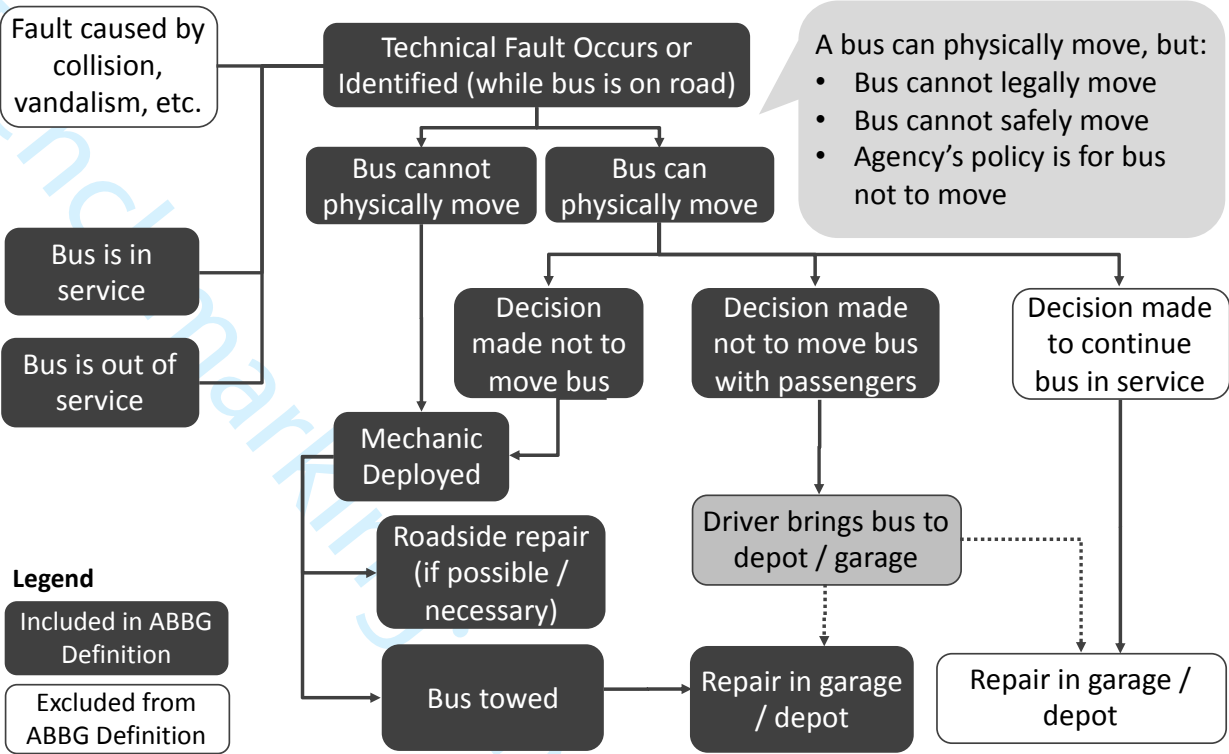


FIGURE 3 Flow Diagram Depicting ABBG Definition of a Road Call due to Technical Faults

The flow diagram in Figure 3 resulted in the following definition for “Number of Road Calls Due to Technical Faults” as agreed by ABBG agencies: The total number of any technical road calls recorded during the year, including road calls due to technical faults that do not directly impact customer service, such as those made during non-revenue service. This includes situations in which the bus is repaired roadside and those in which the bus is driven or towed back to the depot/garage to be repaired. This should include road calls for tire failures, but should exclude "service" road calls, such as those for passenger incidents, collisions, or criminal incidents. Also include change outs due to technical faults, where passengers are moved to another bus, while the bus with a failure returns to the garage (regardless of impact to the passengers).

A common challenge in dealing with data is the possible change in data sources and technology over time. This requires specificity in definition and understanding of when alternative methodologies are being used, documentation of changes, and a balance between preserving meaningful trends and removing previous year data that is no longer comparable. Passenger miles is one data item that often has significant changes in value due to changes in methodology (samples) and technology (APC, customer surveys), but another example is on-time performance. Table 2 lists the drop in performance that occurred for six ABBG organizations when they went from sampling using manual supervisor assessment of on-time performance to data collection using Automatic Vehicle Location (AVL) technology for the entire fleet. As an example, if Agency C previously had reported 90.5% on-time performance using manual supervisor assessment, its performance would have dropped to 82.4% (-8.1%) after

implementation of AVL. The significant impact of AVL on reported performance highlights the need to agencies to manage stakeholder expectations and understanding in advance.

TABLE 2 Drop in On-Time Performance with Implementation of AVL – Trends data from the American Bus Benchmarking Group

Agency	Drop in on-time annual performance after AVL installed (Given as actual change in reported performance)
A	-22%
B	-12.1%
C	-8.1%
D	-8.1%
E	-7.2%
F	-6.9%

Although resource-intensive, ABBG agencies report that the new data items, detailed definitions, and in-depth review of methodologies have led to valuable identification of internal issues and opportunities as well as an improved level of understanding of performance and ways to make changes, as discussed further below under the Data Understanding section.

Annual Standardized Data Collection Methods (2.4, 3.2)

ABBG agencies provide fixed-route data on an annual basis for approximately 130 clearly defined data items. The data is a mix of performance-related data and background data that provide necessary context and understanding. Up to 11 years of data, from 2006 to 2017, were available at the time of this research. Each year, member agencies submit their latest year data as well as any revisions to previous year data that may be necessary due to new methodologies or data systems, the identification of any errors, or the fact that a data item definition has changed per agreement of the group. Thus, the annual collection of data over time is a key contributor to data quality, in providing opportunities for revisions as well as supporting data validation, as discussed below.

Data items are collected from agencies using standardized data sheets, with separate sections for operational, financial, profile, fleet, and staff data. The agreed-upon definitions are included in the data workbooks and any new data items or revised definitions are highlighted when distributed to the ABBG agencies.

One data collection challenge for ABBG is fiscal year definition. Four different fiscal years are used across the ABBG, with fairly good distribution across the groups. During the initial setup of the ABBG, agencies discussed providing calendar year data but agreed that this would be too onerous and result in inaccuracies. Instead, it was agreed that agencies would provide data for their respective fiscal years, as then the benchmarking data is consistent with data that is internally checked and validated. Thus, for those agencies that contributed data to this research, ABBG data were from one of the following time periods (with agencies using that time period indicated):

- July 1 to June 30 (7 agencies)
- October 1 to September 30 (4 agencies)
- January 1 to December 31 (6 agencies)
- April 1 to March 31 (2 agencies)

In-depth Data Validation and Detailed Follow-Up (3.3, 4.5)

The ABBG data validation process, which is an essential piece of ensuring data quality, consists of the following steps, performed for each agency on an annual basis:

- (1) Identification of:
 - a. Missing data items,
 - b. Major changes from previous year data (both in direction and magnitude),
 - c. Internal inconsistencies within that year’s data (for that agency only), and
 - d. Significant differences in performance (direction or magnitude) compared to other agencies.
- (2) Development and issuance of follow-up questions identifying issues and asking for revisions or explanations (if none were provided originally) and subsequent review of responses and additional follow-up as necessary.
- (3) Documentation of explanations whether service-related or methodological.

The repetition of these steps over time adds to data quality and understanding (as discussed below in terms of trend tracking) and is consistent with the iterative, continuous process that is part of the overall ABBG benchmarking approach.

Challenge: Data Understanding

Data understanding in the ABBG is supported by some of the same practices that improve data quality, namely the small group size (2.2), annual cycle (2.4), and in-depth follow-up process (4.5). In addition, there are six sub-elements of Figure 2 that specifically address the challenge of understanding and are discussed in this section.

Bi/multi-dimensional Normalization (4.1)

As mentioned previously and evident in the ABBG fixed-route KPI system (Table 1), nearly all of the KPIs have more than one normalization factor. Trompet *et al.* (2018) provides evidence that conclusions from relative performance differences observed in a single KPI (e.g. only normalized for scale by a single denominator) can be biased. Due to variability between peers (even when all peers are in the United States) of operating characteristics such as commercial speed, trip length, vehicle capacity, and network efficiency (i.e., the amount of deadheading) it is necessary that performance is at least reviewed from two different dimensions, in order to obtain an improved and more balanced understanding of relative performance. This is especially important for bus organizations that have ‘extreme’ values in any of these four identified variable operating characteristics.

In this section, the methodology presented in Trompet *et al.* (2018) to establish variability in service characteristics and understand its implications for performance normalization is applied to 2017 data from up to 19 ABBG agencies for the four operating characteristics identified above as well as two profile characteristics. The results are shown in Table 3, with the characteristics ranked by variability for each of the two categories. The most variable service characteristic by some margin is network efficiency (CV=0.34), defined by the proportion of total vehicle miles that are non-revenue (deadheading) rather than revenue, followed by passenger trip length (CV=0.21). These high variability values highlight the importance of normalizing KPIs in more than one way; for network efficiency, it is important to consider normalizing both per revenue and per total vehicle mile (or per revenue and total vehicle hour), while for trip length, it is important to consider both passenger miles and passenger boardings as normalization factors.

Even with the lower variability ($CV=0.11$) for average commercial speed, the slowest agency (12.1 mph) covers 38% fewer miles per hour than the fastest (19.6 mph), which has a significant impact on relative performance for either of these agencies when a KPI is only normalized by vehicle miles, thus highlighting the importance of also normalizing by vehicle hours. Note that commercial speed is calculated using revenue vehicle hours that exclude layover time (unlike NTD), allowing for a more accurate understanding of in-motion or road speed.

TABLE 3 Variability in Service and Profile Characteristics – American Bus Benchmarking Group 2017 Data

Characteristic	N	μ	Min	Max	σ	CV
<i>Operating Characteristics</i>						
Network efficiency - % of deadheading miles	18	12.1%	5.3%	20.1%	4.1%	0.34
Average passenger trip length (passenger miles per boarding) – miles	19	4.7	3.0	6.5	1.0	0.21
Average commercial speed – mph	19	15.6	12.0	19.0	1.7	0.11
Weighted average vehicle seating capacity	19	38.0	33.5	42.3	2.6	0.07
<i>Profile Characteristics</i>						
Average fleet age	19	8	5.3	11.1	1.7	0.21
Typical hourly driver wage (5 years of experience)	17	\$24.20	\$18.50	\$31.20	\$3.60	0.15

N = Number of bus organizations in sample

μ = Sample average

Min = Minimum value

Max = Maximum value

σ = Standard deviation

CV = Coefficient of variation (σ / μ).

Multi-KPI Analysis (2.2, 4.2)

The ABBG's balanced scorecard framework (Morse *et al.*, 2017; Kaplan and Norton, 1992) and KPI system (see Table 1) is comprehensive and intentionally is not summarized in a composite indexed score, unlike much of the relevant benchmarking literature and exercises (e.g., Al Mamun and Lownes, 2011). This is because the ABBG defines success as good performance across all areas and prioritizes understanding performance differences across these areas. The

ABBG has found that being very good in one area often means performing less well in other areas.

Figure 4 provides a specific example of how, even within one success dimension, a holistic approach combined with bi-dimensional normalization as described above leads to very important conclusions that otherwise would be missed. Figure 4 ranks agencies by their average cost efficiency (cost per output) performance (assessed as: below, the same as, or above the group’s average) for seven operating cost KPIs (the first seven rows), representing total and subcomponent operating costs normalized by either vehicle hour or mile, depending on which one is more appropriate to the cost being measured. Within these cost efficiency KPI rows, one can identify subareas of higher or lower cost for individual agencies. The bottom two rows show cost effectiveness (cost per outcome) performance for total operating cost normalized by boardings and passenger miles. These additional two rows demonstrate that some agencies with high cost efficiency (low cost per output) have low cost effectiveness (high cost per outcome), such as Agency C, while other agencies have low cost efficiency (high cost per output) but high cost effectiveness (low cost per outcome), such as Agency R. In addition, this table provides another example why bi-normalization is important in cases where high variability in operating characteristics exists. Relative cost effectiveness performance for Agency F and Agency Q is very different when normalized per passenger boarding or passenger mile. This is a result of the fact that Agency F has long trip lengths, and Agency Q has short trip lengths. This variability in trip lengths is shown in Table 3.

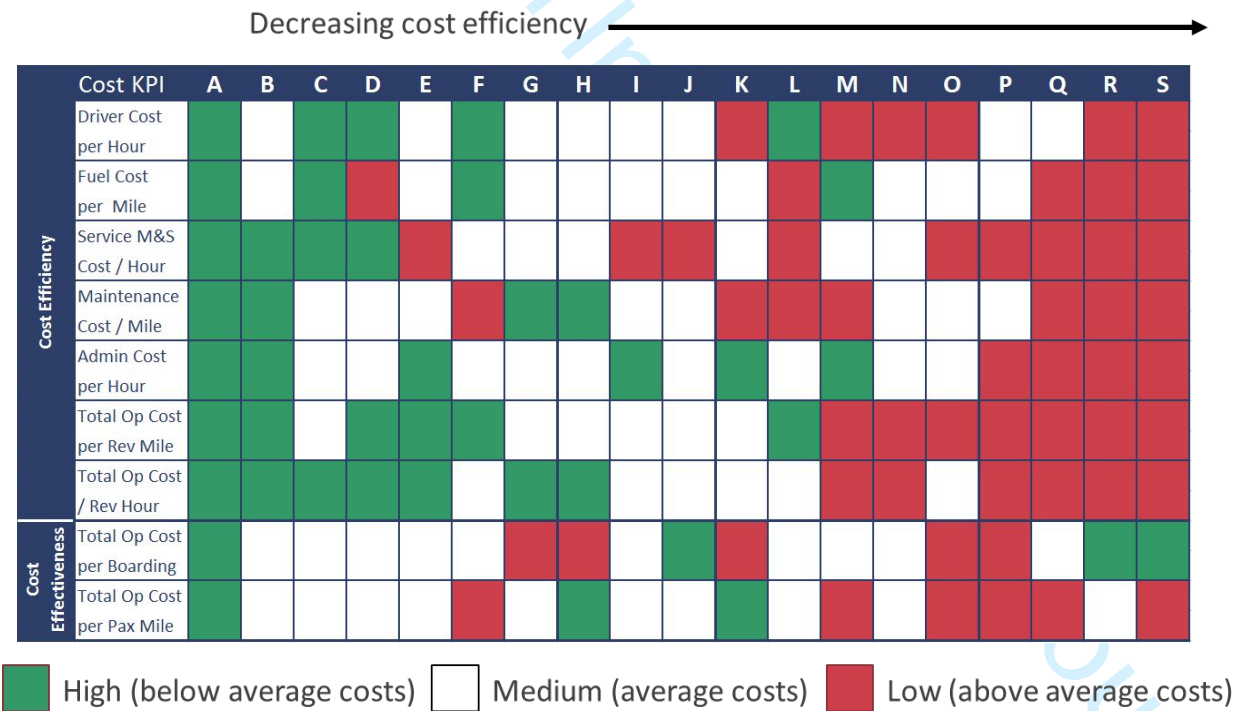


FIGURE 4 A Comparison of Cost Efficiency and Cost Effectiveness Performance – American Bus Benchmarking Group, 2017 Data

Breakdown of Data Items (4.3)

There is a risk when comparing high-level data items of leading to unhelpful or inaccurate generalizations about causation or contributing factors. Of course any further breakdown of data can be resource-intensive so must be worthwhile. The ABBG, as a result of the close working relationship between agencies and the research team, has successfully identified and agreed to a number of worthwhile breakdowns. These include the following:

- (1) 'In-service' vehicle hours broken down into revenue hours and layover hours. This allows for more accurate calculation of revenue speed (excluding layover hours) and also to understand the impact of layover length on driver productivity and on-time performance.
- (2) Fleet utilization broken down into peak fleet utilization, utilization of fleet not in use (maintenance or spare), and fleet utilization by hour of the day (24 hours). The latter in particular allows for an understanding of vehicle and staff utilization, level of service (frequency and coverage), deadheading, and availability of vehicles for maintenance.
- (3) Absenteeism broken down into paid unplanned (primarily sick leave), paid planned (primarily vacation), and unpaid. Paid unplanned leave is the most disruptive to operations but also the most controllable. Paid planned leave depends on state regulations, agency policy, and tenure of staff.
- (4) Vehicle collisions broken down into the categories of preventable, unpreventable, and on-property. Agencies have found the on-property category in particular very helpful to track and benchmark to confirm or identify issues with facilities and training.

Two example KPIs have been included in this paper to demonstrate how the additional breakdown of data can lead to improved understanding of the data. Figure 5 below shows the breakdown of vehicle collision data, with preventable being of most concern, followed by on-property (which are often preventable or at least within the agency's control). Although it is also important to consider collision rates, this graph helps agencies identify where they may have an unusually high proportion of preventable collisions regardless of performance. Thus, even if agency R or S had the lowest collision rates in the group, they may have the greatest potential to reduce their rate further, given the high proportion of preventable collisions (and in the case of agency R, high proportion of on-property collisions). Figure 6 (in the Trend Tracking and Analysis section below) shows the breakdown of absenteeism data for all staff.

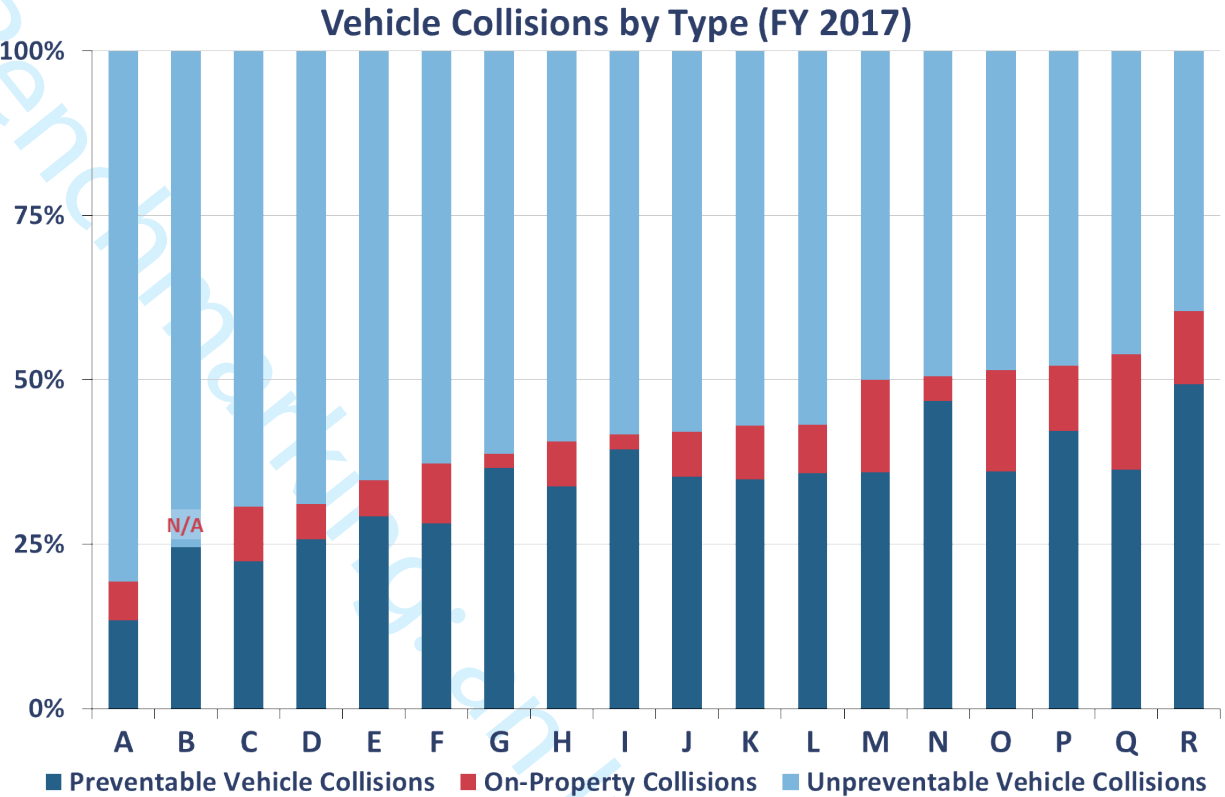


FIGURE 5 Proportion of Collisions due to Preventable, Unpreventable, and On-Property Collisions for the American Bus Benchmarking Group (2017)

Trend Tracking and Analysis (4.4)

Tracking trends across time facilitates data validation as well as the identification of improving or worsening performance. In many cases, other agencies can learn from an agency that is improving in performance regardless of the agency’s relative performance. Trends also inform the agency whether they are going in the right direction or not so they can assess the need for intervention or effectiveness of a change. Trends analysis can also be combined with the breakdown of data into subcomponents. Figure 6 shows one example of this type of analysis from the ABBG, looking at trends in absenteeism for all staff, broken down into the three types as described previously. The ABBG breaks down these data further into staff category (e.g., drivers and maintenance).

In Figure 6, Agency L is fourth highest in total absenteeism but sixth lowest in paid unplanned absenteeism, which has the most direct (and negative) impact on operations. Agency O has faced challenges in paid unplanned absenteeism but has managed to decrease the rate over time. By understanding the information shown in the absenteeism KPI, including the additional information provided by the breakdown of data and tracking trends, ABBG agencies have been able to identify their strengths and weaknesses. They have also been able to share a number of strategies over the years to reduce paid unplanned absenteeism, including providing wellness and fitness facilities and incentives, incorporating absenteeism into performance and promotional reviews, and partnering with unions.

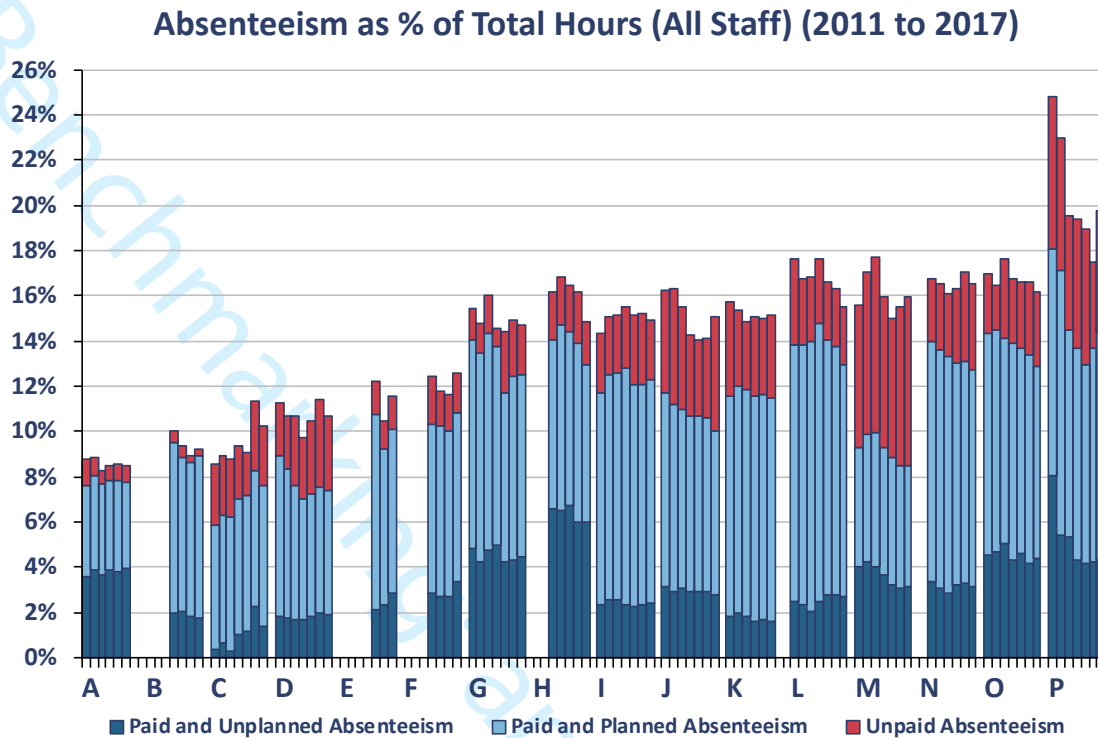


FIGURE 6 Staff Absenteeism Over Time, Broken into Type of Absenteeism for the American Bus Benchmarking Group

Context (2.6)

Understanding relative performance relies on the ability to assess performance within the local and agency context. For the ABBG, differences in state regulatory, economic, and political environments are particularly important to consider given that the ABBG has 24 agencies in 18 states as shown in Figure 1. The ABBG primarily addresses contextual differences through collecting background data, such as the driver wages included in Table 3, and identifying and documenting such differences through agency feedback, the research team's investigations, and small, agency-led topical studies. For example, an ABBG study on absenteeism identified major differences in state Family Medical Leave regulations and agency benefits that were a contributing factor to absenteeism rates (Figure 6).

CONCLUSION

This paper has presented in detail how the ABBG has applied a comprehensive, iterative approach to address the general benchmarking challenges of comparability, data quality, and understanding. Although the application of this research was to mid-sized transit agencies in a U.S. context, it is a model that has been used in other public transport modes and in international groups. Practitioners in public transport and other industries can use this paper as a guide to improve their approaches to peer selection, KPI identification, data collection and analysis, and understanding performance. By covering all the elements presented in the methodology (Figure 2), the ABBG has achieved high-quality, comparable data, leading to improved understanding. As a result, ABBG agencies are able to learn from relative performance, best practices, and lessons learned across the group and apply that learning to their own operations.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

There are several limitations to this research and the framework provided. Benchmarking can be a valuable exercise for agencies in most sectors – but to do it well requires time and commitment. This requirement means that the benchmarking approach described in this paper, as well as any similar research, is most suitably applied to a concise group of agencies (rather than industry-wide). In addition, the data used for this approach are annual and system-wide rather than specific to routes or peak times and thus result in high-level indicators. These indicators consequently require further drill-down analysis by the individual agencies but create a useful starting point. Thus it is strongly advised that participants continuously review the benchmarking results to identify internal actions they can take to improve their performance by applying lessons learned from peers that realized superior performance. Based on ABBG experience, these improvements can take the form of changes in policy, planning, operations, stakeholder communication, and/or program development.

A key research finding and practical implication of this research is that this methodology should not be applied for only for one year, but instead requires a minimum of three years to gain maturity. To ensure relevance over this time and beyond, the KPI system itself (including the definitions) must be continuously reviewed against changing technology, data sources, and priorities, while protecting the comparability of trends where possible. The challenges that this paper focused on (comparability, quality, and understanding) are broad and cross-cutting for all transport and industries. However, it is also important when applying this methodology elsewhere to account for any challenges specific to the chosen participants being benchmarked. In the case of the ABBG, the challenges of the US context needed to be considered, such as different fiscal years, different local and state regulations, and the national requirement for paratransit provision. There is an opportunity for future research to apply this methodology to other industries and in other countries.

The benchmarking methodology used by the ABBG has been designed to be dynamic and continuous. In the near future, this adaptability will be tested both by the opportunity for new metrics that better capture customer impact and by the rise of on-demand services and autonomous vehicles. Future research will be needed to identify how this approach can accommodate these changes and which adjustments will be needed.

REFERENCES

- Al Mamun, M. S., and Lownes, N. E. (2011), "A Composite Index of Public Transit Accessibility", *Journal of Public Transportation*, Vol. 14 No. 2, pp. 69-87. Available at: <http://doi.org/10.5038/2375-0901.14.2.4>.
- Alstete, J. (2008), "Measurement benchmarks or "real" benchmarking? An examination of current perspectives", *Benchmarking: An International Journal*, Vol. 15 No. 2, pp. 178-186. Available at: <https://doi.org/10.1108/14635770810864884>.
- Chu, X., Fielding, G. J., and Lamar, B. W. (1992), "Measuring Transit Performance Using Data Envelopment Analysis", *Transportation Research Part A: Policy and Practice*, pp. 223-230. Available at: [https://doi.org/10.1016/0965-8564\(92\)90033-4](https://doi.org/10.1016/0965-8564(92)90033-4).
- Cook, T. J., Collins, D. G., and Monast, K. (2010), *Benchmarking Guidebook for North Carolina Community Transportation Systems*. North Carolina Department of Transportation. Available at: <https://itre.ncsu.edu/wp-content/uploads/2016/03/FINAL-Short-CT-Guidebook-20110222.pdf>.
- Fielding, G., Babitsky, T., and Brenner, M. (1985), "Performance evaluation for Bus Transit", *Transportation Research Part A, Vol. 19A* No. 2, pp. 73-82. Available at: [https://doi.org/10.1016/0191-2607\(85\)90009-3](https://doi.org/10.1016/0191-2607(85)90009-3).
- Florida Department of Transportation (2014), *Best Practices in Evaluating Transit Performance: Final Report*. Tallahassee, FL: Freight Logistics and Passenger Operations, Transit Office. Available at: https://fdotwww.blob.core.windows.net/sitefinity/docs/default-source/content/transit/pages/bestpracticesinevaluatingtransitperformancefinalreport.pdf?sfvrsn=48878730_0.
- Fu, L., and Xin, Y. (2007), "A New Performance Index for Evaluating Transit Quality of Service", *Journal of Public Transportation*, Vol. 10 No. 3, pp. 47-69. Available at: <http://doi.org/10.5038/2375-0901.10.3.4>.
- Gan, A. F., Gui, F., and Tang, L. (2011) 'System for Transit Performance Analysis Using the National Transit Database', *Journal of Public Transportation*, Vol. 14 No. 3, pp. 87-107. Available at: <http://doi.org/10.5038/2375-0901.14.3.5>.
- Geerlings, H., Klementsitz, R., and Mulley, C. (2006) 'Development of a methodology for benchmarking public transportation organisations: a practical tool based on an industry sound methodology', *Journal of Cleaner Production*, 14, pp.113-123. Available at: <https://doi.org/10.1016/j.jclepro.2005.03.021>.
- Hartgen, D. T., and Horner, M. W. (1997), *Report 163: Comparative Performance of Major U.S. Bus Transit Systems: 1988—1995*. Charlotte, NC. University of North Carolina.
- Hendren, P. (2011), "Bringing Performance-Based Management to Transit". In *Performance Measure of Transportation Systems: Summary of the Fourth International Conference (K. Turnbull, ed.)* (pp. 66-69). Washington, DC: The National Academies Press. Available at: <https://doi.org/10.17226/22535>.
- Henning, T., Essakali, D., and Oeh, J. E. (2011), *A framework for urban transport benchmarking*. Washington DC: World Bank. Available at: <http://www.utbenchmark.in/img/RefDocuments/Home-Ref-1-3-1.pdf>.

Hensher, D., and Daniels, R. (1995), “Productivity Measurement in the Urban Bus Sector”, *Transport Policy*, Vol. 2, pp. 179-195. Available at: [https://doi.org/10.1016/0967-070X\(95\)96748-A](https://doi.org/10.1016/0967-070X(95)96748-A)

Hilmola, O. (2011), “Benchmarking efficiency of public passenger transport in larger cities”, *Benchmarking: An International Journal*, Vol. 18 No. 1, pp. 23-41. Available at: <https://doi.org/10.1108/14635771111109805>.

Hinton, M., Francis, G. and Holloway, J. (2000), "Best practice benchmarking in the UK", *Benchmarking: An International Journal*, Vol. 7 No. 1, pp. 52-61. Available at: <https://doi.org/10.1108/14635770010314954>.

Kaplan, R. S., and Norton, D. P. (1992), “The Balanced Scorecard – Measures that Drive Performance”, *Harvard Business Review*, 7 (1), pp. 71-79.

Karlaftis, M. G. (2004), “DEA approach for evaluating the efficiency and effectiveness of urban transit systems”, *European Journal of Operational Research*, Vol. 152 No. 2, pp. 354-364. Available at: [https://doi.org/10.1016/S0377-2217\(03\)00029-8](https://doi.org/10.1016/S0377-2217(03)00029-8).

Kassens-Noor, E., Kotval-Karamchandani, Z., Brush, A., Doshier, K., and Biskey, M. (2019), “Michigan's public transportation: An application of statewide performance assessment and management”, *Transportation Research Interdisciplinary Perspectives*, 1. Available at: <https://doi.org/10.1016/j.trip.2019.100013>

Kittelsohn & Associates, Inc., Urbitrans Inc., LKC Consulting Services, Inc., MORPACE International Inc., Queensland University of Technology, Y.A. Nakanishi (2003), *The Transit Cooperative Research Program (TCRP). Report 88: A Guidebook for Developing a Transit Performance-Measurement System*. Washington, D.C.: Transportation Research Board. Available at: http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_rpt_88.pdf.

Lema, N., and Price, A. (1995), “Benchmarking: performance improvement towards competitive advantage”, *Journal of Management in Engineering*, 11(1), pp 28-37.

Magd, H. and Curry, A. (2003), “Benchmarking: achieving best value in public-sector organisations”, *Benchmarking: An International Journal*, Vol. 10 No. 3, pp. 261-286. Available at: <https://doi.org/10.1108/14635770310477780>.

Meyer, M. D. (2002), “Measuring System Performance: Key to Establishing Operations as a Core Agency Mission”, *Transportation Research Record*, Vol. 1817 No. 1, pp. 155-162. Available at: <https://doi.org/10.3141/1817-20>

Min, H., and Ahn, Y. (2017), “Dynamic Benchmarking of Mass Transit Systems in the United States Using Data Envelopment Analysis and the Malmquist Productivity Index”, *Journal of Business Logistics*, Vol. 38 No. 1, pp. 55-73. Available at: <https://doi.org/10.1111/jbl.12148>.

Morse, L., Trompet, M., Barron, A., and Graham, D. (2017), “Development of a Key Performance Indicator System to Benchmark Relative Paratransit Performance”, *Transportation Research Record*, pp. 1-8. Available at: <https://doi.org/10.3141/2650-01>.

Perk, V., and Kamp, N. (2004), *Benchmark Rankings for Transit Systems in the United States*. Tallahassee, FL: State of Florida Department of Transportation. doi:NCTR-527-03/FDOT BC137-43. Available at: <https://www.nctr.usf.edu/pdf/527-03.pdf>.

Ryus, P. et al. (2010), *TCRP Report 141: A Methodology for Performance Measurement and Peer Comparison in the Public Transportation Industry*. Washington, D.C.: Transportation Research Board. Available at: <https://doi.org/10.17226/14402>.

Trompet, M., Anderson, R. J., and Graham, D. J. (2009), "Variability in Comparable Performance of Urban Bus Operations", *Transportation Research Record: Journal of the Transportation Research Board*, pp. 177-184. Available at: <https://doi.org/10.3141/2111-20>.

Trompet, M., Anderson, R. J., and Graham, D. J. (2018), "Improved understanding of the relative quality of bus public transit using a balanced approach to performance data normalization", *Transportation Research Part A: Policy and Practice*, Vol. 114, pp. 13-23. Available at: <https://doi.org/10.1016/j.tra.2017.11.019>.

Trompet, M., Parasam, R., and Anderson, R. J. (2013), "Benchmarking Disaggregate Customer Satisfaction Scores of Bus Operators in Different Cities and Countries", *Transportation Research Record*, pp. 14-22. Available at: <https://doi.org/10.3141/2351-02>.

Verbich, D., Badami, M., and El-Geneidy, A. (2017), "Bang for the buck: Toward a rapid assessment of urban public transit from multiple perspectives in North America", *Transport Policy*, Vol. 55, pp. 51-61. Available at: <https://doi.org/10.1016/j.tranpol.2016.12.002>.

Zairi, M. (1994), "Benchmarking: The Best Tool for Measuring Competitiveness", *Benchmarking for Quality Management & Technology*, Vol. 1 No. 1, pp. 11-24. Available at: <https://doi.org/10.1108/14635779410056859>.