Designing New York City Subways' Key Performance Indicators to Improve Service Delivery & Operations

Alla Reddy (corresponding)
Senior Director, System Data & Research (SDR), Operations Planning
New York City Transit
2 Broadway, Office A17.92, New York, N.Y. 10004-2208
Tel: (646) 252-5662; Email: *Alla.Reddy@nyct.com*

Alex Lu
Senior Specialist, Capital Planning & Programming
Metro-North Railroad
345 Madison Ave., Floor 3, New York, N.Y. 10017-3739
Tel: (212) 340-2684; Email: *lexcie@gmail.com*

Mikhail Boguslavsky
Associate Staff Analyst, System Data & Research (SDR), Operations Planning
New York City Transit
2 Broadway, Cubicle A17.110, New York, N.Y. 10004-2208
Tel: (646) 252-5663; Email: *Mikhail.Boguslavsky@nyct.com*

Brian Levine
Associate Staff Analyst, System Data & Research (SDR), Operations Planning
New York City Transit
2 Broadway, Cubicle A17.100, New York, N.Y. 10004-2208
Tel: (646) 252-5541; Email: *Brian.Levine@nyct.com*

Word Count: 245 Words (Abstract) + 5,417 Words (Text) + 6 Figures * 250 Words = 7,162 Words

ABSTRACT

Balanced Scorecard (BSC) is widely used in private industry and the public sector to monitor Key Performance Indicators (KPI) and to help achieve strategic outcomes. This concept is widely used in transit industry carrier-regulator contractual relationships and performance monitoring. After a fact-finding mission to South East Asia, New York City Transit (NYCT) adopted KPIs to continually improve service delivery performance. Subway line-level KPIs based on BSC concepts were introduced in conjunction with a Line General Manager program and numerous incremental performance management initiatives. Following a reorganization recreating functional departments—car equipment, stations, and rapid transit operations—BSC was applied at departmental levels resulting in maintenance-oriented Passenger Environment Survey (PES)-KPI and operations-oriented Service (S)-KPI. Weightings of indicator subcomponents were assigned using surveys of customer priorities. The KPIs provide one number representing overall performance, but they also make it possible to identify each subcomponent's contribution. The KPI design processes also generated public feedback prompting NYCT to tighten underlying performance standards. Today, PES-KPI and S-KPI are reported monthly to the MTA Board's NYC Transit Committee. Advantages of these indicators include high-level visibility and ease of communication, timely report availability, and detailed diagnostics. These factors, together with reinvigorated competitive spirit between divisions triggered by reorganizations, resulted in a much more proactive organization focused on using performance scores to take corrective action. Wait Assessment, the principal component of S-KPI, improved 2.5% on the heavily crowded "1" through "6" lines in 2012 compared to 2011 even as ridership increased steadily systemwide.

INTRODUCTION

Balanced Scorecard (BSC) is widely used in private industry and public sector to monitor Key Performance Indicators (KPI) and achieve strategic outcomes. New York City Transit's (NYCT) accurate operations reporting and performance measurement traditions date back to early 20th century (*1*); performance was normally measured for individual functions or departments, providing high levels of accountability. A survey (*2*) conducted for NYCT's parent agency Metropolitan Transportation Authority (MTA) found more than 300 discrete indicators, but no overarching framework for organizing or combining them into higher-level measures reflecting overall organizational goals. NYCT's industry benchmarking, stakeholder demands, and the burgeoning Line General Manager (LGM) program together led to an effort to design a set of KPIs to measure agency-wide operational performance using BSC frameworks.

Background

Local transit advocates designed a "State of Subways" report card in 1997 (*3*), applying fairly basic weighting methodologies (*4*) to six KPIs. Unfortunately it's "graded on a curve," resulting in relative performance rankings by line, rather than objective measurements of how NYCT has improved.

Transport benchmarking bodies like L'Union Internationale des Transports Publics (UITP) have long collected KPI data (*5*). Community of Metros (CoMET), an international peer organization, proposed KPIs for comparing subway performance between cities (*6*,*7*) in six categories: Innovation, Financial, Internal Processes, Customer, Safety & Security, and Environment. While CoMET KPIs evaluate all performance aspects, their primary purpose is benchmarking, not measuring continuous internal improvements—and are thus not necessarily suitable for public dashboard reporting.

In 2009, NYCT participated in TCRP's South East Asia fact-finding mission entitled "Performance Measurement and Outcomes." BSC and KPI concepts were ubiquitous throughout transit agencies there (*8*), in governing regulator-carrier relationships, performance regimes, and internal benchmarking (*9*). Staff felt that introduction of these management tools could improve NYCT's service delivery performance.

Several agencies are experimenting with region-wide measures to assess transit networks' overall health and performance with KPI-like methodologies. London's focuses on four targeted high-level measures: customer satisfaction, passenger trips, excess journey time, and % scheduled kilometers operated (*10*,*11*). Likewise, Chicago's effort (*12*) didn't weigh each indicator to produce a composite region-wide measure, although ratios of discrete indicators were used to determine high-level effectiveness (*13*). Amtrak's KPI efforts (*14*,*15*) included financial and ridership indicators, but NYCT decided early on that only operational indicators would be included—the KPI program's goal was to achieve operational improvements.

Introduction of a composite KPI using BSC methodologies could produce "one number," publicly reported, as quantitative scorecard of how subway system is performing as a whole. This provides useful monitoring frameworks for senior management, and objective statistics for outside observers.

Literature Review

Balanced Scorecard was essentially a process management strategy introduced in 1987 at Analog Devices (*16*), popularized by Kaplan & Norton (*17*). BSC is a concept, rather than specific methods or indicators; as such, diverse applications are found in the literature. Applications range from providing organizing frameworks for performance measures to being centerpiece of companywide strategic planning (*18*). Initial applications to transportation were in logistics process engineering (*19*), although transit applications soon followed (*20*,*21*).

Early transit agency applications did not gain widespread acceptance, even as BSC flourished elsewhere in transportation service delivery. Typical applications include: assessing trucking quality performance (*22*), supply chain performance (*23*), maintenance performance (*24*), and transport sustainability (*25*).

KPI is a related framework. Whereas BSC provides a single composite quantity encompassing and embracing many qualitative, competing (and sometimes orthogonal) priorities of the company, a KPI-style regime represents each corporate strategic priority as one or more indicators on a "dashboard" without weighting or arbitrating their relative importance.

Monthly KPIs are often presented as long-term (e.g. 12-month) rolling averages, capturing trends while being sensitive to short-term improvements, which show up as smaller changes. 12-month "look-back" periods provide incentive for sustained improvement, while allowing monthly performance checkpoints.

INCREMENTAL PERFORMANCE MEASUREMENT IMPROVEMENTS

Starting 2007, NYCT increased internal reporting frequencies on both service reliability and passenger environment performance, for buses and subways. Improvements realized from these programs made them precursors to the BSC-based KPI in place today.

Daily Temperature Monitoring Program

The "Big Chill" temperature monitoring program was launched in summer 2007. On days when forecast outside temperature exceeded 85º F, traffic checkers surveyed interior temperatures of randomly selected subway cars and buses. The vehicle spec-based passing criterion was temperature below 78º F. Daily results were reported directly to the President, holding operating managers accountable (*26*).

In September, awards were issued to best- and worst-performing bus depots and train yards (*27*). Systemwide, subway cars passing increased from 95.3% in 2007 to 97.4% in 2008, and held steady at 97.2% in 2009. Buses achieved steady improvement over three summers, with passing rates of 95.4% in 2007, 96.9% in 2008, and 97.7% in 2009.

Monthly Litter and Cleanliness Reports Under Station Cleaning Pilot Program

NYCT had an established Passenger Environment Survey (PES) semiannual report, measuring service quality from customers' perspective. Although environmental conditions can be difficult to assess objectively, PES standards were developed from quality control criteria used by operating departments (*28*). In November 2008, PES measured 68 discrete indicators in four categories (cleanliness, customer information, equipment, and operations) across four environments (subway car, stations, bus, and express bus). In all environments, critical indicators are litter and cleanliness conditions, classified into four categories: None, Light, Moderate, and Heavy.

NYCT's Station Cleaning Pilot Initiative hired 313 additional station cleaners (boosting their ranks to 1,430) and modified cleaning assignments so one employee is no longer responsible for a single aspect at multiple stations (*29*). Agreements were reached with cleaners' union requiring employees to perform multiple tasks at "their" station within one single shift, increasing job ownership and accountability for results.

With this program, NYCT issued monthly snapshot litter and cleanliness reports directly to appropriate operating departments (Buses, Car Equipment, and Stations). This provided frequent feedback, allowing rapid management corrective action during quarterly performance assessment periods.

Special assistants in NYCT President's Office monitored these reports; the President occasionally communicated with operating supervisors directly regarding their performance scores.

This increased performance reports' visibility and reinvigorated station management. Although passing standard was either None or Light, managers requested additional tabulations showing results only for the None category, effectively tightening internal standards.

Monthly Governor's Report

Beginning 2008, MTA issued reports to New York State Governor on subsidiary agencies' operating performance (*30*), including subway/bus Wait Assessment (WA) and PES litter/cleanliness scores. Monthly reports used prior six months' data to ensure adequate sample sizes, giving rise to the nascent rolling-average reporting framework. This effort emphasized to operating departments the need to be responsive to performance monitoring, and eventually developed into an MTA-wide operational performance dashboard (*31*).

NYC TRANSIT'S LINE GENERAL MANAGER (LGM) PROGRAM

The LGM program, launched in late 2007, was designed around the idea that there should be a single point of accountability for a train service. It was accompanied by substantial reorganization within NYCT Department of Subways, to clarify lines of responsibilities and improve assignment of maintenance priorities. First LGMs were assigned to independent lines (Flushing and Canarsie) where they functioned as "general manager" and were responsible for every aspect of that line's day-to-day operations (*32*), including track and car maintenance, train operations, dispatching, and shops and yards (*33*).

Some lines, like the Lexington Avenue "Lex" Line, carried 750,000 weekday passengers, as many as entire subway systems in other American cities (e.g. Washington Metropolitan Area Transit Authority (WMATA) with approximately 723,000 weekday passengers (*34*)). Thus it was felt that an LGM could better oversee issues on their "system" instead of having a large division accountable to all 5 million daily passengers.

The LGM program was later rolled out to remainder of the system (Figure 1(c)), with smaller lines being aggregated together (e.g. "B", "Q", and Franklin "S" Shuttle—which were historically the Brooklyn, Flatbush and Coney Island Railway; "J", "Z", and "M" Lines—historically Brooklyn-Manhattan Transit's Eastern Division) coming under auspices of one LGM (*35*). Designed to foster competitive spirit, it was envisioned that friendly rivalry would develop amongst LGMs, who would vie for best performance results, much as lines had done under President's Temperature Study. To this end, monthly Headway Regularity reports showing 12-month moving averages were distributed to LGMs; these reports identified five best- and worst-performing lines, but they focused only on a single aspect of performance.

The management theory behind this type of organizational restructuring harkens back to British Rail's sectorization efforts in 1981, to create strict lines of cost and service accountabilities within otherwise monolithic organizations (*36*). It seeks to harness organizational advantages of vertically integrated private railways (*37*) within the framework of large state enterprises. Similar organizational models were applied to Japan National Railways in the run-up to privatization. NYCT President declared:

> "The change in management structure […] decentralizes the decision-making process, moving the responsibility into the field where managers are responsible for overseeing day to day subway operations. Decisions leading to operational improvements will now be made almost immediately, rather than winding their way along a bureaucratic chain that often took weeks or months." (*38*)

LGMs' selection criteria included broad-based transportation management background, preferably with both maintenance and operations experience. First group of LGMs were selected to ensure at least one seasoned manager from each main operating discipline was within the same "group" of lines, to allow for mutual support. To augment the main program, a management training program was initiated to develop next generation LGMs (*39*).

Line-Level Composite Key Performance Indicator (LLCKPI)

With LGM program's inauguration, NYCT President's Office sought to assess each line's many aspects of operating performance on a "dashboard," an all-encompassing set of key indicators plus composite numeric indices representing overall line performance. In response, subway Line-Level Composite KPI was developed using BSC concepts. "Lines" in LLCKPI coincided with LGM line groupings (Figure 1(b,c)).

To determine indicators' component weights in LLCKPI, NYCT analyzed most recent available Rider Report Cards (RRC) for all lines (*40*,*41*) to determine customer priorities. RRC was a customer satisfaction survey conducted in 2008 and 2009 to get opinions about service reliability and passenger environment on their lines. It also asked customers to rank questions' (i.e. customer concerns') relative importance, which were aggregated among all lines to determine overall priorities. Concerns/issues were converted to related measures already reported: RRC's "reasonable wait times for trains" mapped to NYCT performance metrics Throughput, Wait Assessment, and Headway Regularity. Weights were assigned based on customers' views.

Indicators where line-level measurements were not readily available were omitted from LLCKPI—e.g. "courtesy and helpfulness of station personnel" did not match any existing indicators. Other measures that NYCT reports regularly were not included on RRCs, and therefore omitted—e.g. subway PES reports "door panels out," which impedes customer flow, but it was not flagged as a customer concern.

LLCKPIs were computed using 28 indicators chosen out of hundreds regularly reported (Figure 2(a)). All indicators were quantified as percentages, with actual values measured as ratios of appropriate goals or standards (Figure 2(b)), then converted into points using customer-generated weights to produce LLCKPI scores. Each indicator and combined indices were to be published quarterly.

Notably, proposed LLCKPI combined into a single score for the first time service performance measures (i.e. WA-based, see (*42*)), including operational indicators, and in-vehicle/station passenger environment measures (i.e. PES-based, see (*28*)), including maintenance indicators. These NYCT measures were traditionally assigned to different departments. Combining them was consistent with BSC concepts, and LGM program's stated goal of making one manager responsible for every aspect of the daily commuting experience.

Return to the Traditional Departmental Organizational Structure

The LGM program was always billed as a "pilot program," and NYCT chose to restore former functional and divisional organizational structure in March 2010 (similar to Figure 1(a)). A main issue with the LGM program overlaid on top of the physical subway infrastructure is the presence of shared lines and yards.

Railroads had traditionally dealt with these issues by utilizing concepts of interline ticketing, trackage and haulage rights, and maintenance service agreements. However, this required an entirely new way of thinking within the heavily embedded functional and divisional structure. Additionally, with any reorganization of this order of magnitude, "the most wide-ranging since the creation of the subway system (*38*)," operational performance might be expected to get worse before it got better. Managers and operations personnel needed time to learn and re-learn more effective ways of doing business under the new framework, and systems needed to catch up to provide cost and performance reporting that emphasized new lines of accountabilities.

The final round of LGM reorganizations (Figure 1(c)) recognized realities of NYCT's intertwined infrastructure by naming "Maintenance Group Managers" to oversee shared-track infrastructure issues (*35*). This program was not without its critics, some of whom thought it did not go far enough in decentralizing management control (*43*), whereas others felt that intrinsic infrastructure capacity constraints would challenge even the most responsive possible management teams (*29*). With two years of practical experience, some felt "this scheme had left the line managers bogged down in maintenance calls" (*44*).

While the preceding outlines valid debate on how best to structure a labyrinth of subway lines, ultimately pressures at a senior level, according to Donohue (*45*): "Walder—on the job just one month—had promised to bring new leadership to the MTA," plus a shortage of maintenance resources under severe fiscal constraints (*46*) resulted in LGM program becoming an inconclusive two-and-a-half-year experiment in how best to internally organize large transit operations.

FUNCTIONALLY BASED KEY PERFORMANCE INDICATORS

Addressing return to functional and divisional organizational structure, NYCT President felt that agency-wide applications of BSC concepts were no longer appropriate, and KPIs should be functionally based. However, BSC concepts would be applied within each discipline, and two systemwide KPIs were proposed: PES-KPI for Divisions of Car Equipment and Stations, and Service KPI (S-KPI) for Rapid Transit Operations (RTO). In the reorganization, lines of responsibility in Car Equipment became shop-based (207th St., Coney Island, Corona, etc.); in Stations it became borough-based (geographic management districts—Manhattan, Bronx, etc.); and for RTO it remained line-based (see Figure 1(a)). This allowed each department to be organized differently, yet provided higher-level KPIs for senior management. Comments from stakeholder groups also revealed needs for separate service reliability KPI (*47*), rather than just "one number" for two fundamentally different aspects of transit management.

Passenger Environment Survey KPI (PES-KPI)

PES-KPI was developed using BSC concepts applied to state of maintenance, mostly by compositing existing PES indicators. NYCT's goal was to increase reporting frequency from twice yearly to monthly, effectively expanding and making Governor's Report methodology a routine public reporting item. PES-KPI added indicators beyond litter and cleanliness to monthly reporting.

PES-KPI reports three categories (appearance, equipment, and information) separately for stations and subway cars, summarized from established PES indicators with weights based on customer concerns and management priorities (Figure 3(a)). Overall PES-KPI combines all environment indicators into a single score.

PES-KPI was also introduced for Staten Island Railway (SIR). Although fewer indicators are evaluated under SIR's PES program, they were combined into the same categories for KPI reporting.

Using rolling averages, NYCT achieved increased reporting frequency without additional data collection costs. In-depth PES reports showing individual measures are still published semiannually. PES-KPI became fully effective during second half 2010.

Although LLCKPI framework was readily applicable to maintenance-oriented PES-KPI, for service-oriented S-KPI, more revisions were needed due to the fewer available existing performance measures.

On-Time Performance (OTP) Indicator Revisions

NYCT had used two different service reliability measures of On-Time Performance (OTP): "Absolute OTP" reflecting customer experience, and "Controllable OTP" accounting for only delays under the Authority's control. Although nomenclature differs, this is the basic approach taken by rail operators in Boston with "actual" and "contractual" OTP (*48*); by Network Rail in Great Britain, which measures "external" fraction of delay causes in their public performance measure (PPM); by Amtrak, with "host railroad," "Amtrak," and "third party" delays (*15*); by airlines reporting delays to U.S. DOT (*49*,*50*) in five distinct categories.

Both OTP measures were based on train arrivals at destination terminals. Absolute OTP compared actual arrivals to base schedule for all trips and delays, including delays due to schedule supplements for capital and maintenance work. Controllable OTP compared actual arrivals to daily schedules in effect ("plan of the day" in Britain (*51*)), meaning base schedule with supplements applied; it also included all trips but excluded delays charged to factors such as customer disruptions and police activity. Trains were "on time" if they arrived at destination no later than 5 minutes after scheduled time and didn't skip any scheduled station stops. OTP was measured over 24 hours and was reported separately for weekdays and weekends.

During S-KPI's design process, deficiencies were identified in these standards (Figure 3(b)). "Fixed" Terminal OTP, eventually adopted, combined best elements of Absolute and Controllable measures (*52*). It compares actual arrivals to daily schedules based on service plan in effect, and factors in all delays, including those charged to police and customers. It assesses no penalty for planned platform closures. Major efforts were made by train performance clerks with one-time manual recalculation (reassessing approximately 8,000 train starts per weekday) for data going back one year, to maintain historical continuity by line.

Wait Assessment (WA) Indicator Revisions

NYCT's Wait Assessment (detailed in (*42*,*53*)) measures service reliability en-route, capturing passenger perspectives by defining maximum acceptable wait between actual departures. Actual intervals between trains (headways) are compared to headways defined by schedules in effect.

Responding to stakeholder requests during S-KPI design (*47*), WA was modified to make it stricter overall. Standard for passing intervals was tightened to +25% of scheduled headway from +2/+4 minute (peak/off-peak) values. This reduced bias against infrequent lines, where absolute numbers represented smaller percentages of scheduled headways than for lines with more frequent service. Still, the new +25% threshold is more stringent for most routes (Figure 3(c)). When WA was recalculated for 2009 using +25% threshold, the systemwide indicator decreased from the upper 80% range to upper 70% (though trends over time remained similar). Calculations showed +2/+4 minute standards were approximately equivalent to +43% threshold systemwide, though figures obviously varied from line to line, and period to period. Historical continuity was retained by recalculating existing electronic data.

Although OTP was measured on weekends, WA was not. To address this discrepancy, NYCT implemented weekend WA using same standards (+25% threshold) as weekdays. Automatically collected Automated Train Supervision (ATS-A) data on "1"–"6" lines allowed weekend WA to be reported immediately for those lines (*53*). Sampling methodology was used for BMT and IND divisions, the "7" Line, and Grand Central "S" Shuttle. Division-level results became available after six months of data collection, and line-level results followed after six more months.

NYCT considered using Average Passenger Wait Time (APWT), an alternative to WA. APWT assumes random passenger arrivals, and measures estimated actual passenger wait time compared against scheduled. It reflects the experience per passenger, not per train, and is similar to Excess Wait Time (EWT) described in Trompet et al. (*54*). Although historical data could have been computed, this indicator was deemed too difficult to explain to America's traveling public, particularly with constantly varying headways scheduled by NYCT. Long tails on typical NYCT wait-time distributions also gave rise to averages that seem implausibly low compared to some customers' experience.

Technically, APWT required precise knowledge of unlinked incident ridership (*55*) at each station by time of day (to at least quarter-hourly averages), line, and direction. Although linked ridership data is obtainable from NYCT's Automated Fare Collection (AFC) system, large station complexes like Times Sq.-Port Authority represented a significant directional and line-level ridership allocation challenge. Moreover, determining transfer volumes based on point-of-entry swipe data in

New York's highly interconnected and redundant network (with multiple routing permutations) is no trivial task (*56*). Although this data could have been obtained from a passenger flow model (as is done in London and Hong Kong (*9*)), it was felt that performance measures should be based on physically measureable and publicly verifiable metrics, rather than estimates or projections from mathematical models.

Service Key Performance Indicator (S-KPI)

Applying BSC concepts to service reliability indicators alone posed some difficulty, due to limited number of candidate components. Service PIs published in the Transit Committee meeting agenda were OTP, WA, Peak Throughput, Counts and Causes of Delays, and Mean Distance Between Failures (MDBF). NYCT selected OTP, WA, and MDBF for inclusion in S-KPI—Delay Counts and Peak Throughput were not easily incorporated into percentage scores that S-KPI required, and OTP and WA were thought to sufficiently capture customer experience.

Initially proposed weights were 50% for WA, and 25% each for OTP and MDBF, but they were later revised to 60% WA, 30% OTP, and 10% MDBF. WA was given highest weight because it best reflects customers' perception and experience of service provided; MDBF's weight was reduced because of its inherent limitations.

Though MDBF is one measure of car reliability, for S-KPI purposes it's not a useful management tool. More appropriate statistics might be failures per mile compared to standard specification, and should perhaps account for age and utilization patterns of vehicle fleets. NYCT considered numerous methods attempting to reflect these factors (Figure 4(a)), but those options were not selected due to difficulties in explaining to the public fleets' intrinsic mechanical differences by line, and NYCT's budgetary processes for internal line-level performance goal determination—which are subject to technical and political considerations both internal and external. Figure 4(b) highlights instructive comments received during MDBF options' review process.

Mechanical failures affect OTP and WA adversely, so other S-KPI components are already sensitive to MDBF changes. Additionally, line-level MDBF expressed as percentage of systemwide goal (Figure 4(a), Option 2) fluctuates more than other S-KPI components, particularly as monthly in-service failures approach single digits on some lines. By weighing it more heavily, S-KPI would be tracking MDBF and not overall performance. Figure 4(c) shows some remarks received.

OTP and WA are high-level functional performance measures reflecting customers' view of system performance (*42*). It is thus difficult to imagine an S-KPI that did not heavily rely on those two components. For the public, the combination of these two indicators with MDBF presents an overall picture of factors affecting service.

Internally, S-KPI allows operational management to better understand how indicators relate to each other and plan accordingly. However, data mining to determine how each lower-level, operational measure (throughput, mechanical reliability, terminal delays, customer incidents, crew availability, track work and track outages, etc.) impacts higher-level indicators might have been more helpful. Counts and Causes of Delays are valuable for this reason, but they do not present a single number in an easy-to-comprehend format.

Although WA could potentially be a proxy for other indicators like throughput—if scheduled throughput is not met, WA would likely pick up gaps in service—the two measures don't behave exactly the same. If loss of 3 trains per hour (tph) results in flexing from 4-minute scheduled headways (15 tph) to 5-minute headways (12 tph), all 5-minute intervals would still pass WA at +25%.

While current incarnation of S-KPI leaves room for improvement, it represents nonetheless valiant attempts to apply BSC methodologies to some very tricky performance measurement issues.

WEIGHTING DETERMINATION USING MARKET RESEARCH SURVEY

Before component weights in S-KPI and PES-KPI were finalized, stakeholder input was that an actual passenger perception survey was needed.

NYCT hired an outside firm to conduct market research surveys (*57*), with objectives of assigning weights (or utilities) to service attributes that corresponded to existing indicators. Though subway KPI development motivated the survey, NYCT also included questions about buses, collecting data opportunistically for future bus KPI efforts. 1,046 phone interviews (540 for bus, 506 for subway) were completed in August 2010, with customers who used the system in preceding 30 days (*58*). Most sampled phone numbers were landlines selected through random digit dialing (*59*). A smaller cell phone-only random sample was reached, representing 6% of interviews.

Survey respondents were presented with pairs of attributes and asked to choose which item was most important to them. This was repeated until all attribute pairs (39 bus, 36 subway) were traded off, in patterns reminiscent of round-robin sports tournaments. To compute absolute rather than relative utilities, a multiple paired comparison analysis was conducted, producing unique utilities for each attribute.

This study design was different from RRC surveys, where effectively individual surveys were conducted by line. Random sampling avoided potential self-selection bias, unlike RRC where customers were asked for their opinions but completed surveys only if they wanted to.

Market Research Results

Market research weightings for PES-KPI didn't deviate greatly from original proposals (equipment condition, information, and appearance categories at 45%/35%/20% versus 33%/33%/33%). Customers shifted priorities primarily from appearance to condition. NYCT decided to shift weightings using an average of original and survey weightings, partially reflecting customer concerns.

Several indicators weighted higher by customers were given less weight as their historical performance had consistently been 99% to 100%. If they were given too much weight, PES-KPI would fluctuate very little. Litter and cleanliness weights were reduced only slightly on subway cars even though customers viewed them as less important because they have historically been effective management tools for determining cleaner performance; when they are not properly monitored in stations and on subway cars, overall conditions deteriorate.

For S-KPI, survey respondents weighted MDBF significantly higher than initial proposals (37% versus 10%), apparently due to the MDBF description phrasing: "Having fewer breakdowns and delays due to the mechanical failure of subway cars." Customers correctly viewed MDBF as measuring equipment reliability. However, for reasons already discussed, MDBF should not be a major S-KPI component. Thus NYCT chose to keep the weightings initially proposed. This decision was due to mathematical behavior of MDBF rather than any suggestion that trainset reliability is not important.

PERFORMANCE RESULTS

In September 2010, NYCT S-KPI, PES-KPI, and SIR PES-KPI were added to Transit Committee meeting agenda (*60*), presented in line charts showing 12-month moving averages (Figure 5(a,b)). Revised OTP and WA measures replaced old indicators, and weekend WA was introduced.

PES-KPI Operational Impacts

PES-KPI proved effective in identifying areas where improvements are needed. In 2011, Bronx Station PES-KPI dropped one percentage point (Figure 5(c)), and NYCT received complaints from local politicians (*61*). KPI allowed management to drill down and determine that the Information subcomponent was bringing down PES-KPI scores, and see that the System Map indicator was bringing down Information scores.

For stations to pass System Map standards, two current maps are required: one each in paid and unpaid areas. Many Bronx stations are elevated; on those platforms, relatively few map frames were in place where system maps could be posted. Having identified root causes of poor performance, management installed new map frames. Even based on 12-month rolling average, PES-KPI rose sharply. By December 2011, Information scores had reversed their decline, and Bronx's overall PES-KPI rose along with them.

Wait Assessment-Ridership Correlation

An NYCT study (*42*) demonstrated that subway Headway Regularity had a strong negative correlation with ridership on a line-by-line basis for the year spanning May 2007 to April 2008. Headway Regularity was later replaced as a measure of service reliability by WA, now S-KPI's highest-weighted component (60%). While causal relationships should not be inferred from observational data, factors associated with higher ridership are likely to make it more difficult to maintain even spacing between trains. Larger volumes of passengers contribute to increased platform crowding at busy stations, where trains incur longer dwell times. Higher overall ridership also adds opportunities for disruptive passenger behaviors, like door-holding.

The correlation between WA and estimated ridership by line was similar for the complete years 2008 and 2010 (Figure 6(a,b)). However, the relationship weakened from 2010 to 2012, as measured both by the flatter slope of the fitted regression line and a lower R^2 value (Figure 6(b,c)). Introduction of S-KPI in 2010 coincided with adoption of automatically collected track-occupancy data for the A Division (*53*), enabling daily WA reporting for "1"–"6" lines and Grand Central "S" Shuttle based on 100% data. Also, ATS-A data was later used to revise schedules for the "1" and "6" lines. These developments were complementary, with S-KPI enabling management to identify priorities and respond to problems quickly.

Correlation data suggest management more effectively mitigated potential service reliability issues associated with increased ridership in 2011 and 2012. Systemwide, S-KPI remained stable through 2011 and 2012 (Figure 6(d)), even as ridership increased steadily and several lines continue to operate at their maximum track capacity. WA improved for 15 out of 20 lines in 2012 compared to 2011, including all of the heavily crowded "1"–"6" lines where service regularity had been most difficult to maintain. On average, WA increased 2.5% in 2012 over 2011 for the "1"–"6" lines, based on 100% (not sampled) data. Systemwide gains from WA were offset by declines in OTP (and MDBF) over the same period, though, resulting in stable S-KPI.

LESSONS LEARNED

NYCT's performance management regime underwent dramatic changes several times in recent years. Automated data sources were added, but definitions, reporting frequencies, and presentation of PIs also saw substantial revamp, together with underlying organizational structure and management accountabilities. Below is a summary of our experience:

- **Performance Measures Should Reflect Underlying Organizational Responsibilities:** Performance measures have a duality of functions—when tracked over time, they measure an organization's progress in continuous improvement; when drilled down as snapshots, they identify organizational responsibilities and contributions to overall performance. During tumultuous periods covered by this paper, NYCT was undergoing fundamental changes in responsibilities, accountabilities, and philosophy behind how subway systems should be organized. Frequent changes in performance metrics reflect senior management's desire to manage the system differently and introduce incentives that reward performance improvements within specific organizational frameworks. While this may be confusing to outside oversight authorities, performance measures generally should reflect organizational responsibilities while being aligned with customer goals.

- **Historical Continuity Is Important When Changing Performance Measures:** If performance measures are changed to reflect new organizational structures, it's important to measure impacts of organizational changes on performance. It is thus essential to maintain historical continuity by applying new performance standards retroactively to suitable "lookback" periods, enabling organization's forward progress to be compared against history. However, historical continuity should not be used as a reason to avoid changing established performance measures to account for new organizational direction. Instead, raw operations data could be used to recalculate higher-level performance measures and recreate performance history.
- **Performance Feedback Should Provide Opportunities for Short-Term Improvements:** To incentivize and achieve improvements, performance feedback should be provided sufficiently frequently to allow short-term corrective actions to have meaningful impacts on long-run scoring. This is not "cheating"; the point of measurements is providing benchmarks and incentivizing progress. Rather than seeing performance indicators as an audit function, they should be an integral part of the operational management toolkit, without compromising independent authority of the performance measurement czar.
- **Many Factors Affect Selection of KPI Weights:** KPI was introduced, then its constituent components were changed and re-weighted multiple times. Organizational priorities, management needs, customer perspectives, and indicator variability all contributed to ongoing conversations regarding what and how much should be included in KPIs. Every organization should work through similar performance planning processes in determining appropriate KPIs for their organizational and strategic vision. As organizations evolve, KPI components and weightings should also change.
- **Balanced Scorecard Methods Can Shape Management Focus:** KPIs don't allow operational management to ignore any indicators that comprise them, because reduction in any subcomponent will reduce overall KPI. It also allows explicit tradeoffs between different indicators in resource-constrained scenarios, driven by weightings set by policy. These properties make KPIs a good high-level management metric for assessing organizational performance.
- **Balanced Scorecard Methods Can Highlight Issues in Existing Performance Indicators:** NYCT's KPI design process provided impetus and opportunity to revise and improve existing component indicators. This is an added benefit of BSC methods.
- **Competitive Dynamics Can Re-energize Operations Management:** A key consequence of NYCT's LGM program and BSC application is it re-introduced friendly rivalry between divisions and challenged operational management to find ways to continually improve performance, even long after the LGM program's demise. Internal benchmarking can be helpful in this way provided performance data remains independent and objective.
- **Performance Measures' Ease of Communication Is Critical:** Both for public at large and for service delivery management, performance measures that are intrinsically hard to understand or relate to are not necessarily helpful. Although some measures might have interesting mathematical properties that provide correct management incentives, if supervisors can't relate to them and they are viewed as a black box, it will be difficult to focus operational improvement efforts. If customers can't relate to measures, they may be seen as meaningless "scores" that doesn't represent their viewpoint, or worse still, they may challenge the methods' validity.

ACKNOWLEDGEMENTS

The authors would like to thank Steven Aievoli, Glenn Lunden, Larry Gould, Peter Cafiero, Shoshana Cooper, H. Robert Menhard, Anthony Cramer, John Cucarese, P.J. Diskin, Mindy Rhindress, Stan Hsieh, Brittany Dutchman, Peter Harris (MTA), ALGM Dan Mazzella, LGM John Hoban, LGM Herbie Lambert, LGM Louis Brusati, LGM Joseph Ragusa, LGM Paul McPhee, former NYCT President Howard H. Roberts, former NYCT President Thomas F. Prendergast, Atty. Gene Russianoff, and Dr. Jan Wells for providing feedback during development of NYCT's Key Performance Indicator program. All errors or omissions are responsibility of the authors. Opinions expressed are the authors' and do not necessarily reflect official policy or positions of New York City Transit, Metro-North Railroad, the Metropolitan Transportation Authority of the State of New York, or any of its constituent operating agencies.

REFERENCES

(1) New York City Board of Transportation. *Transit Record*—Volume I (1921).
(2) Newhouser, Robert J. *An Assessment of Performance Measures*—Metropolitan Transportation Authority, New York (2003).
(3) Straphangers Campaign. *State of Subway Report Card*—New York Public Interest Research Group (1999). http://www.straphangers.org/statesubs99/statesubs99.html. Retrieved on March 7, 2014.
(4) Straphangers Campaign. *How We Developed Our Profiles and MetroCard Ratings*. http://www.straphangers.org/statesub12/BigMeth_Web.pdf. Retrieved on March 7, 2014.
(5) L'Union Internationale des Transports Publics. *Millennium Cities Database for Sustainable Mobility* CD-ROM (2001).
(6) Anderson, Richard. Improving the Management and Delivery of Mass Public Transportation in Cities. http://www3.imperial.ac.uk/pls/portallive/docs/1/8595698.PDF. Retrieved on March 7, 2014.
(7) London Underground. *TfL Rail and Underground Annual Benchmarking Report*. http://www.tfl.gov.uk/assets/downloads/tfl-rail-and-underground-benchmarking-report-2012.pdf. Retrieved on March 7, 2014.
(8) Harrington-Hughes, Kathryn. *TCRP Research Results Digest 95: International Transit Studies Program Report on the Spring 2009 Mission – Performance Measurement and Outcomes*.
(9) Reddy, Alla V., A. Lu, and T. Wang. Subway Productivity, Profitability, and Performance: A Tale of Five Cities. *Transportation Research Record* 2143 (2010), pp. 48-58.
(10) London Underground. *Performance Report*. http://www.tfl.gov.uk/corporate/modesoftransport/londonunderground/1592.aspx. Retrieved on March 7, 2014.
(11) Mayor of London. *London Underground Performance Reports*. http://data.london.gov.uk/datastore/package/london-underground-performance-reports. Retrieved on March 7, 2014.
(12) Chicago Transit Authority. Performance Metrics, March 2013. http://www.transitchicago.com/assets/1/performance_metrics/Performance_Metrics_-_Mar_2013.pdf. Retrieved on March 7, 2014.
(13) Gallucci, Grace and J.G. Allen. Regional Transit Performance Measures at Chicago's Regional Transportation Authority, TRB Paper #11-3362.
(14) Scovel, Calvin L. III. Inspector General's Report on Amtrak's Actual and Forecast Financial Performance. http://www.oig.dot.gov/sites/dot/files/Amtrak_Semiannual_5-17-10.pdf. Retrieved on March 7, 2014.

(15) National Rail Passenger Corporation. Amtrak Monthly Performance Report, March 2013. *http://www.amtrak.com/ccurl/1009/146/Amtrak-Monthly-Performance-Report-March-2013.pdf*. Retrieved on March 7, 2014.

(16) Schneiderman, Arthur M. *Analog Devices: 1986-1992, The First Balanced Scorecard.* Boxford, Mass. (1999). *http://www.schneiderman.com/Concepts/The_First_Balanced_Scorecard/BSC_INTRO_AND_CONTENTS.htm*. Retrieved on March 7, 2014.

(17) Kaplan, Robert S. and D.P. Norton. *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business School Press, Boston, Mass. (1996).

(18) Rohm, Howard. *Using the Balanced Scorecard to Align Your Organization*. Balanced Scorecard Institute, January 2008, Cary, N.C.

(19) Liberatore, Matthew J. and T.C. Miller. A Framework for Integrating Activity-Based Costing and the Balanced Scorecard into the Logistics Strategy Development and Monitoring Process. *Journal of Business Logistics*, Vol. 19, No. 2 (1998), pp. 131-154.

(20) Walker, A. and L. Aft. Linking Strategic and Operational Performance Using the Balanced Scorecard. *1999 APTA Bus Conference*, Cleveland, Ohio, 2~6 May, 1999, pp. 34-43.

(21) Phillips, J.K. An Application of the Balanced Scorecard to Public Transit System Performance Assessment. *Transportation Journal*, Vol. 43 (2004), Issue 1, pp. 26-55.

(22) Poli, P.M. and C.A. Scheraga. A Balanced Scorecard Framework for Assessing LTL Motor Carrier Quality Performance. *Journal of the Transportation Research Forum*, Volume 57, Issue 3 (2003), pp. 105-130.

(23) Brewer, P.C. and T.W. Speh. Using the Balanced Scorecard to Measure Supply Chain Performance. *Journal of Business Logistics*, Vol. 21, No. 1 (2000), pp. 75-93.

(24) Ray, David. Four Leadership Principles to Help Transportation Professionals (MMC2006-001). *Transportation Research E-Circular* #098, pp. 154-167.

(25) Pei, Y.L., A.A. Amekudzi, M.D. Meyer, E.M. Barrella, and C.L. Ross. Performance Measurement Frameworks and Development of Effective Sustainable Transport Strategies and Indicators. *Transportation Research Record* 2163 (2010), pp. 73-80.

(26) Donohue, Pete. NYC Transit reports that most subway cars are well air-conditioned. *New York Daily News*, August 3, 2008. *http://www.nydailynews.com/new-york/nyc-transit-reports-subway-cars-air-conditioned-article-1.312604*. Retrieved on March 7, 2014.

(27) Roberts, H.H. and D. Mazzella. *Feel the Heat (or not) Awards Luncheon—Program*, New York City Transit Authority, July 27 (2007).

(28) Lu, Alex, S. Aievoli, J. Ackroyd, H.C. Carlin, and A.V. Reddy. Passenger Environment Survey: Representing the Customer Perspective in Quality Control. *Transportation Research Record* 2112 (2009), pp. 93-103.

(29) Stagl, Jeff. New York City Transit's Road to Reorganization. In *Progressive Railroading*. *http://www.progressiverailroading.com/passenger_rail/article/New-York-City-Transits-Road-to-Reorganization--15342*. Retrieved on March 7, 2014.

(30) New York Public Authorities Law, Article V, Title 11, Section 1269(d)—Submission of Strategic Operation Plan. *http://www.lawserver.com/law/state/new-york/ny-laws/ny_public_authorities_law_1269-d*. Retrieved on March 7, 2014.

(31) Metropolitan Transportation Authority. PERS *Operating Performance Dashboard*. *http://www.mta.info/persdashboard/performance14.html*. Retrieved on March 7, 2014.

(32) New York City Transit. *MTA NYC Transit Names Line General Managers for 7 and L*, Press Release, December 6, 2007. *http://new.mta.info/mta-nyc-transit-names-line-general-managers-7-and-l*. Retrieved on March 7, 2014.

(33) Orlando, Barbara. Subways' Line General Manager Program on Track. In *At Your Service*, December 2008-January 2009, New York City Transit Employee Newsletter.

(34) Washington Metropolitan Area Transit Authority. Transit Ridership Trends and Markets, March 2009. http://www.wmata.com/pdfs/planning/FINAL%20Transit%20Ridership%20and%20Market%20Trends%20Report.pdf. Retrieved on March 7, 2014.
(35) New York City Transit. *MTA NYC Transit Names Group and Line General Managers For All Subway Lines*, Press Release, August 7, 2009. http://new.mta.info/mta-nyc-transit-names-group-and-line-general-managers-all-subway-lines. Retrieved on March 7, 2014.
(36) Joy, Stewart. Public and Private Railways. *Journal of Transport Economics and Policy*, Vol. 32, Part 1 (1997), pp. 27-49.
(37) Allen, J.G. and A. Lu. Organizational Regimes for Commuter Rail: Improving Train Service Worldwide. *Transportation Research Record* 2146 (2010), pp. 100-108.
(38) New York City Transit. *MTA NYC Transit Introduces First Group of Subway Line Managers*, Press Release, October 27, 2008. http://new.mta.info/mta-nyc-transit-introduces-first-group-subway-line-managers. Retrieved on March 7, 2014.
(39) Orlando, Barbara. Expanded Line General Managers Program Aims for Line-by-Line Teamwork Across Subway System. In *At Your Service*, August-September 2009, New York City Transit Employee Newsletter.
(40) Sander, E.G. and H.H. Roberts. Letters to NYC Transit Subway Riders on the 2nd Rider Report Card. http://enterprise.mtanyct.info/survey/english/letters_new.asp. Retrieved on March 7, 2014.
(41) New York City Transit. Rider Report Card Goes Express Again! Riders To Grade NYC Transit & MTA Bus Express Service City Wide. Press Release, August 11, 2009. http://new.mta.info/rider-report-card-goes-express-again-riders-grade-nyc-transit-mta-bus-express-service-city-wide. Retrieved on March 7, 2014.
(42) Cramer, Anthony N., J. Cucarese, M. Tran, A. Lu, and A.V. Reddy. Performance Measurements on Mass Transit. *Transportation Research Record* 2111 (2009), pp. 125-138.
(43) Savas, E.S. Slow Train to Better Service. *New York Times*, Op-Ed, December 16, 2007. http://www.nytimes.com/2007/12/16/opinion/nyregionopinions/16CIsavas.html. Retrieved on March 7, 2014.
(44) Kabak, Benjamin. In Subway Shake-up, Line Managers Lose Some Oversight. http://secondavenuesagas.com/2010/02/26/in-subway-shake-up-line-managers-lose-some-oversight/. Retrieved on March 7, 2014.
(45) Donohue, Pete. NYC Transit President Howard Roberts resigns in wake of Jay Walder's elevation to MTA chair. *New York Daily News*, November 4, 2009. http://www.nydailynews.com/new-york/nyc-transit-president-howard-roberts-resigns-wake-jay-walder-elevation-mta-chair-article-1.413476. Retrieved on March 7, 2014.
(46) New York City Transit. *MTA New York City Transit Restructures Department of Subways*, Press Release, February 26, 2010. http://new.mta.info/mta-new-york-city-transit-restructures-department-subways. Retrieved on March 7, 2014.
(47) Permanent Citizens' Advisory Council. Minutes Matter, a Review of Performance Metrics at the MTA, January 2011. http://www.pcac.org/wp-content/uploads/2011/01/minutes-matter-12711.pdf. Retrieved on March 7, 2014.
(48) Massachusetts Bay Commuter Railroad. *Track Record: Performance Report Card*. http://www.mbcr.net/Report_Card/TrackRecord_North_Apr13.pdf. Retrieved on July 30, 2013.
(49) Bureau of Transportation Statistics. Airline On-Time Performance and Causes of Flight Delays. https://www.rita.dot.gov/bts/help_with_data/aviation/index.html. Retrieved on March 7, 2014.
(50) U.S. DOT. Reporting the Causes of Airline Delays and Cancellations, Docket No. OST 2000-8164, Final Rule. *Federal Register*, Vol. 67, No. 227 (November 25, 2002), pp. 70535-70545.

(51) Office of Rail Regulation. *Rail Performance*—Public Performance Measure up to and including 2010-11 Q3. *http://webarchive.nationalarchives.gov.uk/20110601212617/rail-reg.gov.uk/upload/pdf/nrt-ch2-railperformance.pdf*. Retrieved on March 7, 2014.

(52) New York City Transit Authority. *Implementation of Modified Train Performance Indicators and Introduction of Key Performance Indicators (KPI)*—Presentation to the MTA Board's NYC Transit Committee, September 21, 2010.

(53) Levine, B., A. Lu, and A.V. Reddy. Measurement of Subway Service Performance at New York City Transit: Case Study with Automated Train Supervision Track-Occupancy Data. *Transportation Research Record* 2353 (2013), pp 57-68, TRB Paper #13-2997.

(54) Trompet, M., X. Liu, and D.J. Graham. Development of Key Performance Indicator to Compare Regularity of Service Between Urban Bus Operators. *Transportation Research Record* 2216 (2011), pp. 33-41.

(55) Frumin, Michael and J. Zhao. Analyzing Passenger Incidence Behavior in Heterogeneous Transit Services Using Smartcard Data and Schedule-Based Assignment. *Transportation Research Record* 2274 (2012), pp. 52–60.

(56) Reddy, Alla V., A. Lu, S. Kumar, V. Bashmakov, and S. Rudenko. Entry-Only Automated Fare-Collection System Data Used to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. *Transportation Research Record* 2110 (2009), pp. 128-136.

(57) Rhindress, Mindy. *Design Considerations for the Customer Priorities Study*—Abt SRBI, July 28, 2010.

(58) Rhindress, Mindy, S. Hsieh, and B. Dutchman. *2010 Customer Priorities Survey Final Quadrant Chart Results*—Abt SRBI, October 19, 2010.

(59) Rhindress, Mindy and S. Hsieh. *2010 Customer Priorities Technical Memorandum*—Abt SRBI, September 16, 2010.

(60) New York City Transit Authority. *Transit Committee Meeting Agenda, September 2010*. *http://www.mta.info/mta/news/books/*. Retrieved on March 7, 2014.

(61) Beekman, Daniel. MTA plans to fix Bronx subway stations, the city's worst-kept lines. *New York Daily News*, July 27, 2011. *http://www.nydailynews.com/new-york/bronx/mta-plans-fix-bronx-subway-stations-city-worst-kept-lines-article-1.159350*. Retrieved on March 7, 2014.

LIST OF FIGURES

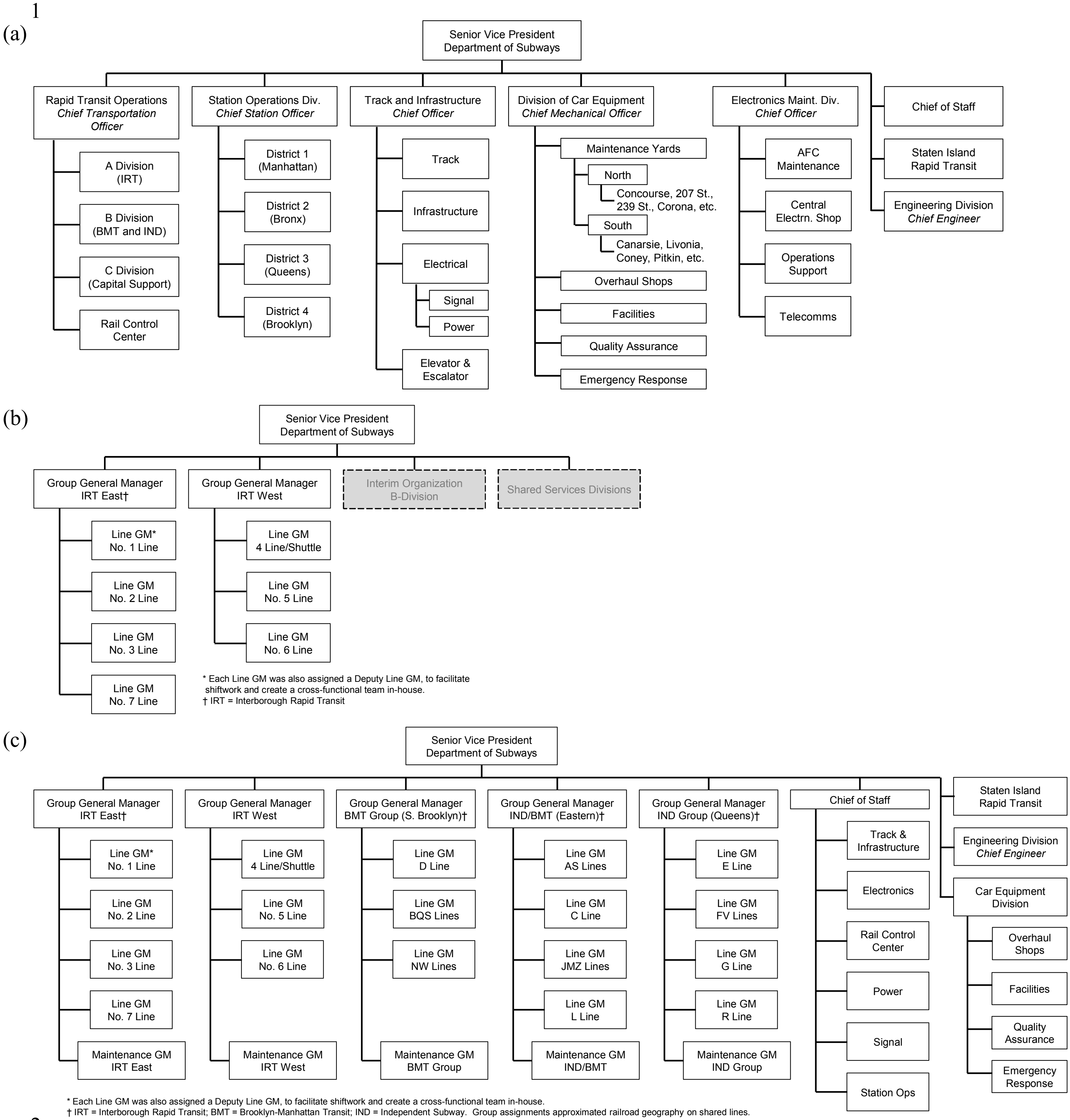


FIGURE 1. Diagrammatic Representation of NYCT Department of Subways Organizational Structure: (a) Traditional Functionally Based Structure in Effect Prior to the Line General Manager Program; (b) Interim Structure During LGM Pilot Program; (c) Structure After Full Implementation of LGM Program.

(a)

Customer Concern	Rider Reported Importance (5 = Highest)	Related Line-Level NYCT Performance Measures	Concern Weight in LLCKPI	Indicator Weight in LLCKPI	Actual "A" Line Performance in First Half 2009	LLCKPI Points Contributed on "A" Line
1. Reasonable wait times for trains	5	(1) Scheduled peak throughput vs. guideline (2) Headway Regularity (18 hour) (3) Wait Assessment (18 hour)	19%	6.3% 6.3% 6.3%	98% 84% 70%	6.2 5.3 4.4
2. Minimal delays during trips	5	(1) Controllable OTP (24 hour) (2) Absolute OTP (24 hour) (3) Failure rate/MDBF	17%	5.7% 5.7% 5.7%	69% 71% 83%	3.9 4.0 4.7
3. Adequate room on board at rush hour	5	(1) Peak loads and cordon counts (off-peak)	13%	13%	100%	13.0
4. Station announcements that are easy to hear	5	Due to ongoing installation of public announcement equipment, not yet available in all stations	—	—	—	—
5. Cleanliness of stations	4	(1) Station litter conditions (pre-AM peak) (2) Station litter conditions (post-AM peak) (3) Station cleanliness (pre-AM peak) (4) Station cleanliness (post-AM peak)	10%	2.5% 2.5% 2.5% 2.5%	74% 64% 85% 83%	1.9 1.6 2.1 2.1
6. Train announcements that are easy to hear	4	PES train announcement	10%	10%	90%	9.0
7. Cleanliness of subway cars	3	(1) Subway car litter (pre-AM peak) (2) Subway car litter (post-AM peak) (3) Subway car cleanliness (pre-AM peak) (4) Subway car cleanliness (post-AM peak)	7%	1.8% 1.8% 1.8% 1.8%	95% 93% 95% 93%	1.7 1.6 1.7 1.6
8. Sense of security in stations	3	Measured by MTA Police, but not available by line	—	—	—	—
9. Sense of security on trains	3		—	—	—	—
10. Working elevators and escalators in stations	2	PES functioning escalator/elevator	5%	5%	91%	4.6
11. Comfortable temperature in subway cars	2	PES functioning heat and air conditioning	5%	5%	100%	5.0
12. Lack of scratchitti in subway cars	1	PES subway car scratchitti	2%	2%	62%	1.2
13. Lack of graffiti in subway cars	1	(1) Cars with no interior graffiti (at terminal) (2) Cars with no exterior graffiti (3) Cars with no graffitied windows	2%	0.7% 0.7% 0.7%	96% 100% 95%	0.7 0.7 0.7
14. Lack of graffiti in stations	1	PES station graffiti	2%	2%	100%	2.0
15. Signs in stations that help riders find their way	1	(1) PES system maps correct (2) PES station map availability	2%	1% 1%	30% 89%	0.3 0.9
16. Signs in subway cars that help riders find their way	1	PES subway car maps	2%	2%	99%	2.0
17. Availability of MetroCard Vending Machines	1	PES functioning MVM/MEM in stations	2%	2%	89%	1.8
18. Ease of use of subway turnstiles	1	PES functioning turnstiles in stations	2%	2%	100%	2.0
19. Courtesy and helpfulness of station personnel	1	Not measurable by NYCT	—	—	—	—
20. Train announcements that are informative	1	Assessed in PES train announcements above	—	—	—	—
		Total	**100%**	**100%**	—	**86.9**

(b) Formula for Converting MDBF (Numerical Indicator) to a Percentage of Goal in Customer Concern #2, Indicator #3:

MDBF% = (MDBF by Line ÷ Annual Systemwide MDBF Goal), subject to a maximum of 100%

Calculation Example ("A" Line): MDBF points = 120,350 miles ÷ 145,000 miles × 5.7% (weighting) = 83% × 5.7% = 4.7 points out of 5.7 maximum

FIGURE 2. Initial NYCT Line-Level Composite Key Performance Indicators Using First Half 2009 Data: (a) Worked Example for the "A" Line—207th St. Inwood, Manhattan to Far Rockaway, Queens; (b) Mean-Distance Between Failures (MDBF) Computation Detail.

(a)

Environment		Stations			Subway Cars	
Category	Overall Wt.	Indicators	Component Wt.	Overall Wt.	Indicators	Component Wt.
Appearance	37%	Litter	15%	34%	Litter	13%
		Cleanliness	15%		Cleanliness	13%
		Graffiti	7%		Graffiti	4%
					Windows	4%
Equipment	31%	Escalators/Elevators	14%	33%	Climate Control	16%
		Fare Vending Machines	11%		Lighting	9%
		Booth Microphone	3%		Door Panels	8%
		Turnstiles	3%			
Information	32%	System Maps	9%	33%	Announcements	11%
		Map Available	9%		Destination Signs	10%
		Psgr. Info. Centre	10%		System Maps	9%
		Uniform	4%		Uniform	3%

(b)

- Terminal OTP may not reflect customer experience, because many customers board or disembark at stations en-route and not at termini.
- Actions taken to improve OTP do not necessarily improve customer service:
  - Adding scheduled recovery time before terminals would not improve performance en-route—and may lead to “terminal holdouts” when trains are held outside terminal interlockings waiting for platform docking;
  - No penalties were assessed for trains departing early en-route;
  - On capacity-constrained corridors, over-emphasis of OTP could encourage reduction in scheduled service to improve statistics.
- Absolute OTP penalized long-term schedule changes for construction.
  - In November 2009, Controllable OTP was 97.3% for the “B” line while Absolute OTP was 4.7%, because stations were bypassed between Newkirk Avenue and Kings Highway during the Brighton Line Station Rehabilitation project. Some observers found disparities between these two measures confusing.

These measures did not distinguish between incident delays (non-controllable) and incident recovery (which is very much controllable). External incidents had only minor impact on Controllable OTP under these standards.

(c)

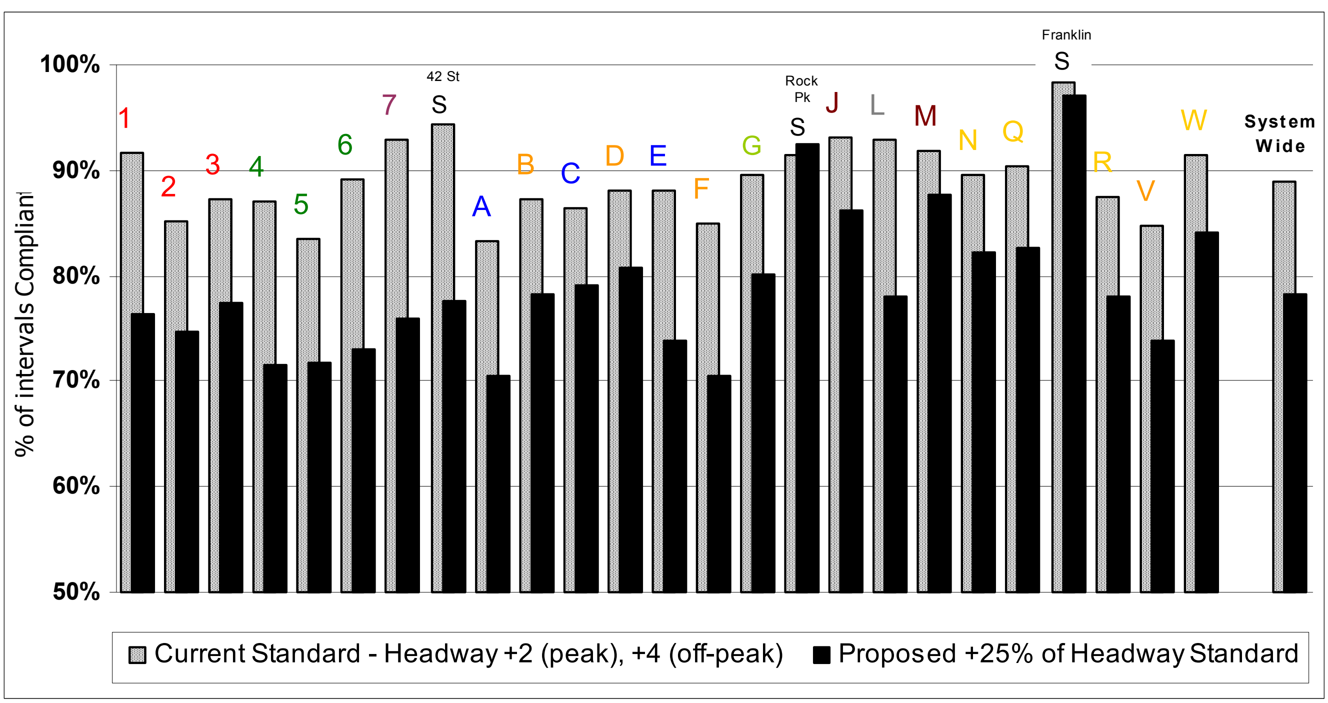


FIGURE 3. Development Process for Functionally Based Performance Indicators: (a) Current PES-KPI Weighting Factors; (b) Identified Deficiencies in the Terminal OTP Performance Measure; (c) Wait Assessment by Line, Old versus Proposed Standards.

(a)

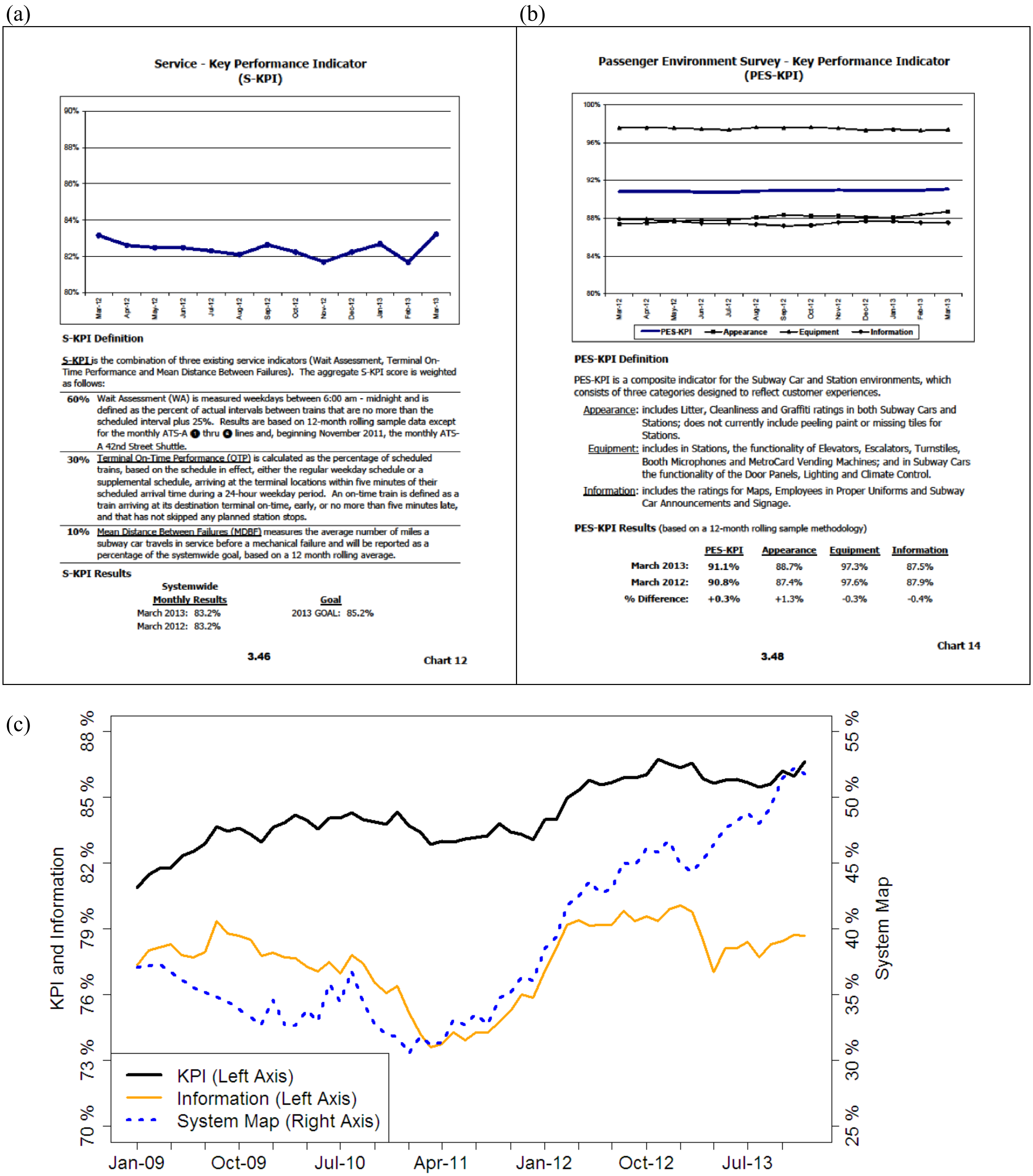
Assess Period	Line	Feb 08-Jan 09 MDBF	Option 1 Goal By Line GOAL	Option 1 Goal By Line % of Goal	Option 2 Systemwide Goal GOAL	Option 2 Systemwide Goal % of Goal	Option 3 SKPI MDBF Std by Line Feb 07-Jan 08 STD	Option 3 SKPI MDBF Std by Line % of Std
Jan-09	1	107,104	130,000	82.4%	145,000	73.9%	99,020	100.0%
Jan-09	2	232,471	200,000	100.0%	145,000	100.0%	256,227	90.7%
Jan-09	3	131,726	145,000	90.8%	145,000	90.8%	142,698	92.3%
Jan-09	4	184,577	160,000	100.0%	145,000	100.0%	186,946	98.7%
Jan-09	5	216,773	200,000	100.0%	145,000	100.0%	199,158	100.0%
Jan-09	6	123,157	150,000	82.1%	145,000	84.9%	139,696	88.2%
Jan-09	7	165,012	130,000	100.0%	145,000	100.0%	163,035	100.0%
Jan-09	A	27,993	200,000	14.0%	145,000	19.3%	51,251	54.6%
Jan-09	B	113,382	125,000	90.7%	145,000	78.2%	132,482	85.6%
Jan-09	C	159,534	250,000	63.8%	145,000	100.0%	170,363	93.6%
Jan-09	D	61,154	250,000	24.5%	145,000	42.2%	75,423	81.1%
Jan-09	E	229,100	250,000	91.6%	145,000	100.0%	213,760	100.0%
Jan-09	F	110,465	80,000	100.0%	145,000	76.2%	115,786	95.4%
Jan-09	FS	116,935	60,000	100.0%	145,000	80.6%	101,103	100.0%
Jan-09	G	57,531	80,000	71.9%	145,000	39.7%	53,840	100.0%
Jan-09	GS	56,088	25,000	100.0%	145,000	38.7%	67,575	83.0%
Jan-09	H	102,196	250,000	40.9%	145,000	70.5%	106,133	96.3%
Jan-09	JZ	138,438	200,000	69.2%	145,000	95.5%	158,525	87.3%
Jan-09	L	212,067	75,000	100.0%	145,000	100.0%	188,547	100.0%
Jan-09	M	166,834	200,000	83.4%	145,000	100.0%	160,548	100.0%
Jan-09	N	328,185	140,000	100.0%	145,000	100.0%	330,407	99.3%
Jan-09	Q	159,534	125,000	100.0%	145,000	100.0%	179,101	89.1%
Jan-09	R	76,530	80,000	95.7%	145,000	52.8%	90,117	84.9%
Jan-09	V	59,506	250,000	23.8%	145,000	41.0%	41,279	100.0%
Jan-09	W	105,713	100,000	100.0%	145,000	72.9%	125,794	84.0%
Jan-09	SYS	135,318	145,000	93.3%	145,000	93.3%	128,724	100.0%

- **Option 1:** Line-level MDBF result (weighted based on actual total failures in service by line divided by car mileages charged to line) as a fraction of Car Equipment Department's internal MDBF fleet-level goals by equipment type (blended based on fleet allocation plan), subject to a maximum of 100%.
- **Option 2:** Line-level MDBF result as a fraction of publicly reported systemwide MDBF goal (max. 100%).
- **Option 3:** Line-level MDBF result as a fraction of prior year's line-level MDBF, plus 10%. Year-over-year improvement of 10% is NYCT's internal benchmark commonly used to set departmental goals.

(b) "Option 3 can be perceived as overly complex, almost opaque."
"Option 2 penalizes lines assigned old cars, and boosts scores for lines with newer cars. I don't think this is a negative from the audience's perspective. This is a savvy audience who would be surprised if new cars didn't score better. And, they are especially on the lookout for new cars which perform poorly."
"In Option 3, the 'standard' actually came out with a 100% factor systemwide, even though systemwide MDBF average was less than goal. That's a peculiar standard, and we'd be open to criticism for grading on a curve."

(c) "Let's say average MDBF is 100k, and say lines that use R-160 cars achieve 200k. With MDBF getting 25% of weight, you would be unreasonably hiding poor performance on WA or OTP."
"As MDBF has risen to very high levels, the addition of just a very small few (one or two in some cases) delays attributable to car failures can significantly lower numerical MDBF performance. The same thing is true for OTP as you approach 95% and even higher. [… an arithmetical illustration] helped educate customers to actual nature of sensitivity of the statistic and, most importantly, just how fantastic (and fragile) ultra-high MDBF and OTP performance is."

FIGURE 4. Development Process for the MDBF Indicator Component of S-KPI: (a) Options for Converting from Raw MDBF to Percentage Score; (b) Initial Stakeholder Comments Regarding Proposed Options; (c) Feedback Regarding MDBF Weight Factors.



S-KPI Definition

S-KPI is the combination of three existing service indicators (Wait Assessment, Terminal On-Time Performance and Mean Distance Between Failures). The aggregate S-KPI score is weighted as follows:

60%	Wait Assessment (WA) is measured weekdays between 6:00 am - midnight and is defined as the percent of actual intervals between trains that are no more than the scheduled interval plus 25%. Results are based on 12-month rolling sample data except for the monthly ATS-A ❶ thru ❻ lines and, beginning November 2011, the monthly ATS-A 42nd Street Shuttle.
30%	Terminal On-Time Performance (OTP) is calculated as the percentage of scheduled trains, based on the schedule in effect, either the regular weekday schedule or a supplemental schedule, arriving at the terminal locations within five minutes of their scheduled arrival time during a 24-hour weekday period. An on-time train is defined as a train arriving at its destination terminal on-time, early, or no more than five minutes late, and that has not skipped any planned station stops.
10%	Mean Distance Between Failures (MDBF) measures the average number of miles a subway car travels in service before a mechanical failure and will be reported as a percentage of the systemwide goal, based on a 12 month rolling average.

S-KPI Results

Systemwide Monthly Results	Goal
March 2013: 83.2%	2013 GOAL: 85.2%
March 2012: 83.2%	

3.46 **Chart 12**

PES-KPI Definition

PES-KPI is a composite indicator for the Subway Car and Station environments, which consists of three categories designed to reflect customer experiences.

Appearance: includes Litter, Cleanliness and Graffiti ratings in both Subway Cars and Stations; does not currently include peeling paint or missing tiles for Stations.

Equipment: includes in Stations, the functionality of Elevators, Escalators, Turnstiles, Booth Microphones and MetroCard Vending Machines; and in Subway Cars the functionality of the Door Panels, Lighting and Climate Control.

Information: includes the ratings for Maps, Employees in Proper Uniforms and Subway Car Announcements and Signage.

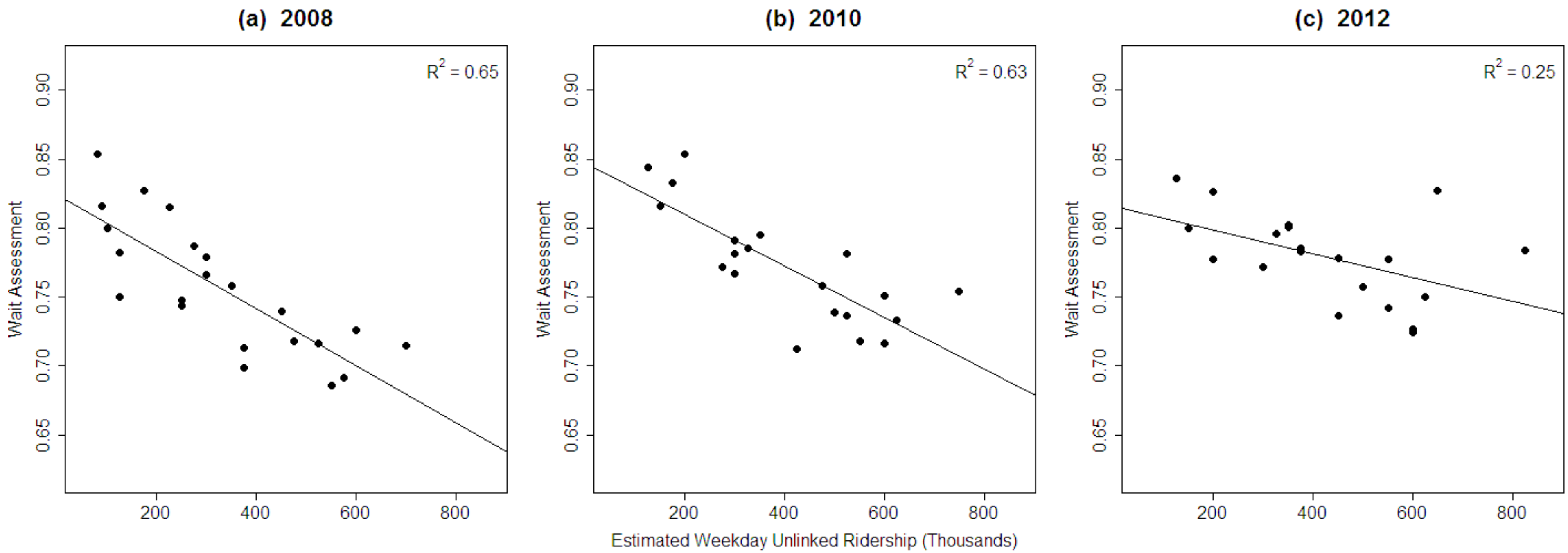
PES-KPI Results (based on a 12-month rolling sample methodology)

	PES-KPI	Appearance	Equipment	Information
March 2013:	**91.1%**	88.7%	97.3%	87.5%
March 2012:	**90.8%**	87.4%	97.6%	87.9%
% Difference:	**+0.3%**	+1.3%	-0.3%	-0.4%

Chart 14

3.48

1 **FIGURE 5.** Application of New York City Transit's KPI and its Subcomponents: (a) New York City
2 Transit Committee Agenda Reports from June 2013 Showing S-KPI Results; (b) PES-KPI Results; (c)
3 Bronx Station PES-KPI, and its Information and Station Map Subcomponents.



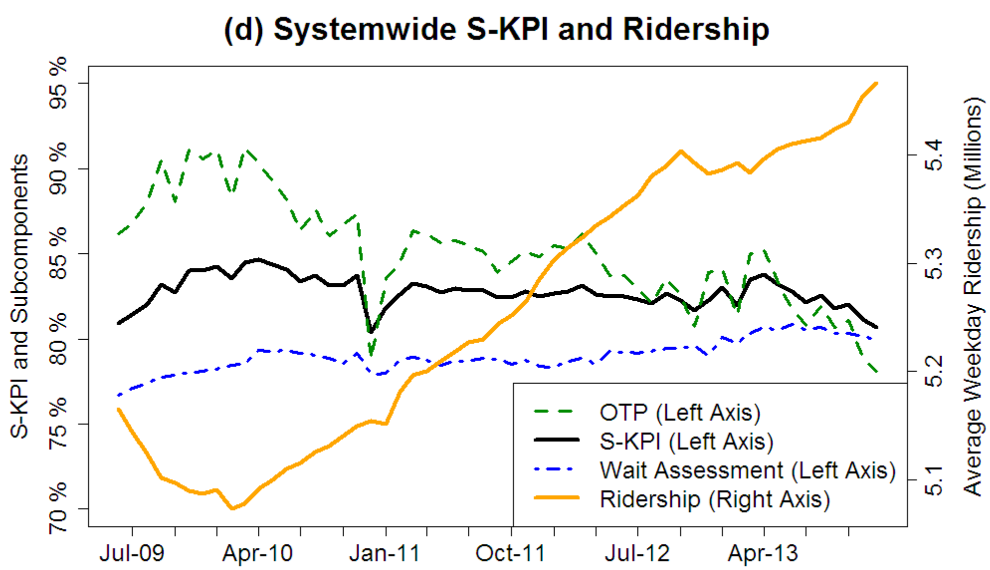


FIGURE 6. Relationships Between Ridership and Wait Assessment: (a)-(c) Weakening Correlation Between Estimated Unlinked Ridership and Wait Assessment from 2010 to 2012; (d) S-KPI and its Subcomponents, Compared to Systemwide Ridership.

Note: (d) Monthly data is used for the OTP subcomponent of S-KPI, while Wait Assessment is based on monthly data for the A Division and 12-month moving average data for the B Division—resulting in a smoother Wait Assessment series compared to OTP. The OTP series reflects a large impact from December 26-27, 2010 blizzard. Ridership data is presented as a 12-month moving average to remove seasonal effects.