The Standardization of Track Gauge on North American Railways, 1830 - 1890

Early North American railways chose different track gauges partly on the basis of differing engineering traditions and partly for mutual compatibility. The resulting dynamic process produced nine distinct common-gauge regions by the 1860s. Growing demand for interregional traffic and increasing cooperation among railways yielded incentives to resolve this diversity, and the specific regional pattern of gauges led to selection of 4'8.5" as the continental standard. The case offers support for aspects of differing views on the role of path dependence in determining features of the economy.

During the fifty years after 1830, six different track gauges came into widespread use on North American (U.S. and Canadian) railways, and several others were used occasionally. At the height of diversity during the 1860s, local standard gauges predominated in nine different regions, but "breaks of gauge" were encountered on interregional routes. By late in the century, nearly all this diversity was resolved, and 4 feet 8 1/2 inches (hereafter 4'8.5"; 1435 mm.) became the standard gauge of a substantially integrated continental railway network.

Track gauge is simply the width between the inside faces of a pair of rails. Standardization of gauge facilitates the exchange of rolling stock, enabling freight shipments and passenger traffic to pass over the track of multiple companies. Under a great variety of historical industry structures, local railways have found it profitable to participate in this through traffic. Thus, the emergence and persistence of diversity of gauge call for an explanation, as does the timing of this diversity's resolution.

The most complete explanation that has been given for this process is that of George R. Taylor and Irene Neu.[1] They note that most of the earliest North American railways were built either to serve strictly local transport needs or else to further the rival ambitions of different commercial centers to capture the trade of the hinterland. As a consequence, early railway builders saw little reason to adopt a common gauge. Taylor and Neu explain the later resolution of diversity in gauge as the result both of an increase in demand for interregional transport and of the takeover of railways by financial capitalists who sought to maximize the value of the railway companies rather than the commerce of particular cities.

Taylor and Neu's explanation is essentially correct as far as it goes, but it leaves certain features of the history unexplained: How did particular gauges come to dominate particular regions—and why did some lines use idiosyncratic gauges? How did markets, other institutions, and various sorts of agents facilitate or hinder the resolution of diversity? Why, ultimately, did 4'8.5" rather than another gauge emerge as the continental standard?

To address these questions, it is necessary to look in micro-level detail at the dynamics of the gauge-selection process. In doing so, it becomes evident that the process was path dependent, in that later outcomes depended on the specific course of preceding events rather than simply on such *a priori* factors as technology, tastes, and factor endowments.

The case of track gauge thus offers an opportunity to test and add empirical content to the theoretical arguments that are made regarding path dependence—both the favorable views of Paul A. David and W. Brian Arthur and the severe qualifications to the relevance of path dependence suggested in the joint work of S.J. Liebowitz and Stephen E. Margolis.[2] The case offers support for aspects of both views and may point the way to a fruitful synthesis.

In view of these arguments, it is worth making clear at the outset that this paper does not focus on whether the historical process has given us the "wrong" standard gauge. In a narrow sense it apparently did, in that most (but not all) railway engineers today would find a broader gauge technically and economically superior, although only to a small extent.[3] But this paper does not assert that early choices for 4'8.5" were wrong given the conditions of their time or that markets, other institutions, or entrepreneurs "failed" in not converting since. The lesson that some economists have drawn from the work of David and Arthur, and the view to which Liebowitz and Margolis respond most vigorously, is that path dependence is fundamentally a source of market failure through lock-in to a suboptimal technique. David and Arthur's definitions, however, emphasize that path dependence has to do fundamentally with the way in which a process of allocation evolves.[4] And although both, especially Arthur, highlight relative inefficiency as a potential consequence of path dependence, even this inefficiency does not necessarily involve market failure. We return to this point in the conclusion.

The emphasis here in this paper is on how allocation evolved—that is, on how regional standard gauges and then a continental standard emerged. The mechanism proves to combine elements of a model proposed by Arthur with types of behavior proposed by Liebowitz and Margolis as an alternative to Arthur's model. The paper also gives attention to the extent to which this mechanism was systematically optimizing as markets, other institutions, and profit-seeking agents promoted or hindered cost-saving standardization.

To understand a path-dependent process, it is necessary to consider not only summary statistics but also the individual events that in turn affected subsequent events. The paper begins with a high-level overview of the history of gauge in North America, introducing features of the process to be explained. It then addresses in some detail the dynamics of emerging diversity, followed by the resolution of this diversity.

Overview of the Gauge Selection Process

Commercial railways in North America began in 1830-32 as several simultaneously built lines introduced four different gauges—4'8.5", 4'9" (1448 mm.), 4'10" (1473 mm.), and 5'0" (1524 mm.). During 1838-48, a second wave of innovation introduced the substantially broader gauges of 5'4" (1626 mm.), 5'6" (1676 mm.), and 6'0" (1829 mm.). Beginning in 1871, a third wave of experimentation brought further diversity in the form of narrow gauges, primarily 3'6" (1067 mm.) in Canada and 3'0" (914 mm.) in the United States.

The gauge of 4'8.5" was always adopted by more new railways than any other, but from the late 1830s to the early 1860s the proportion of new mileage built to that gauge steadily dropped from 87 percent to 44 percent (figure 1). This decline was due both to the introduction of broad gauges and to the relatively faster growth of regional networks using 4'10" and 5'0". As a result, the 4'8.5" gauge's share of mileage in service declined over the same period from 80 to 55 percent (figure 2).[5]

[Figures 1, 2 here]

By the mid-1860s, the eastern half of North America was divided into nine major regions defined by prevalent gauges—that is, by regional standards: (1) eastern Canada and part of Maine, 5'6"; (2) New England and most of New York, 4'8.5"; (3) southern New York, parts of neighboring states, and a route west to the Mississippi, 6'0"; (4) the greater part of New

Jersey, 4’10”; (5) the mid-Atlantic region, from Pennsylvania south to North Carolina, 4’8.5”; (6) the South, east of the Mississippi, except for North Carolina and part of Virginia, 5’0”; (7) most of Ohio and parts of neighboring states, 4’10”; (8) the Midwest (west of Ohio), 4’8.5”; and (9) the trans-Mississippi South, 5’6”.

The mid-1860s represented the height of diversity not only as measured by the overall shares of gauges (figure 2), but also as measured by an index of continental standardization:

Year:	1850	1855	1860	1865	1870	1875	1880	1885	1890
Index:	24	17	18	18	42	55	65	68	90

This index, essentially a Herfindahl-Hirschman concentration index normalized to a scale of 100, gives the average percentage of the total continental railway route accessible from each originating point without crossing a break of gauge.[6] Thus, from 1855 to 1865, an average of 17 or 18 percent of the entire North American network could be reached on a common-gauge route from each point on the network. The index is a measure of the degree of network integration, insofar as integration depends on gauge.[7]

The resolution of gauge diversity began in the late 1860s with the formation of links among the three previously separate 4’8.5”-gauge regions. Primarily for this reason, the continental standardization index more than doubled to 42 in 1870. Thereafter, this index continued to increase due both to the renewed predominance of 4’8.5” in new construction (figure 1) and to conversions of established lines to that gauge.

The pattern of gauge conversion also changed markedly in the late 1860s. From 1830 through 1864, scattered lines totaling about 300 miles of track were converted. During 1865-1869 alone, well over 1,000 miles converted. Thereafter entire regions converted, amounting to over 10,000 miles of track during the 1870s and 20,000 miles during the 1880s.

The first common-gauge regions converted were those using 5’6”, 4’10”, and 6’0”. At the same time, total 5’0”-gauge route actually more than doubled between 1865 and 1885, due both to new construction in the South and to conversion of nearly 1,000 miles of railways that had used 4’8.5”, 4’10”, and 5’6”. Nevertheless, more than 1,000 miles in the South converted to 4’8.5” in 1881 and 1882, followed by the remaining 15,000 miles in 1885 and 1886.

Even as this long-standing diversity in gauge was being resolved, a “narrow gauge fever” from 1871 to 1883 brought a new wave of diversity (figures 1, 2). New narrow-gauge construction dropped dramatically after 1883, and mileage in service more slowly.

The Dynamics of Emerging Diversity

The emergence of diversity of gauge was conditioned by two sorts of incentives for each local railway line: preferences for specific gauges and an interest in adopting a common gauge in order to exchange traffic. What sort of aggregate gauge-selection process should we expect to result from this combination? Arthur considers a similar problem in a non-spatial context.[8] In his model, agents vary in their preferences for specific techniques, but their interest in using a common technique overrides their specific preferences once the difference in numbers of adoptions passes a critical value. As a result, the allocation process begins evolving randomly, depending on the specific preferences of early adopters, but it eventually "locks in" to one or another technique.

Railways differ from the agents in Arthur's model in that spatial relationships matter: they have an incentive to use the same gauge as neighboring lines, not necessarily the majority gauge of the system as a whole. To consider how allocation evolves in such a context, I developed a spatial analog to Arthur's model, modeling railway lines as the cells of a lattice choosing gauges on the basis both of their specific preferences and of the prior (or expected) choices of neighbors and of these neighbors' common-gauge networks. As the model is analytically intractable, I "solved" it numerically using computer simulation. The process typically results in a high degree of regional standardization together with diversity at the level of the whole "continent." Moreover, the specific outcome of the process proves quite sensitive to specific events that can be characterized as random: the gauge preferences of particular early lines and the specific sequence of construction.

There is no space here for a detailed explication of the model, but its essential logic is straightforward.[9] The first railway line built in a region chooses its gauge simply on the basis of its preference. Subsequent lines soon give more weight to the value of common-gauge connections. Thus a common-gauge region expands around the first line built. Eventually this region comes into contact with other common-gauge regions, merging with those that happen to have chosen the same gauge and "competing" for further expansion with regions of different gauges. The eventual distribution of gauges tends to be quite asymmetrical and to vary substantially from the initial distribution of gauge preferences.

Liebowitz and Margolis point out that the results of Arthur's model depend on agents not exercising foresight and not coordinating their choices.[10] The same is true for my model. If new railway lines foresee the expansion of distant common-gauge networks and if they influence each others' choices, the result is greater initial standardization. As we shall see, in historical fact foresight and coordination were limited, particularly in the earlier years. But as these factors grew over time, they substantially affected the gauge selection process.

We consider, next, the origins of engineering traditions favoring various gauges and, then, how the two sorts of incentives affected the process of diffusion of these gauges. We then examine how the mechanism reflected in my model, as well as foresight and coordination, led to the emergence of regional standard gauges.

Sources of Variation: The Engineering Traditions

The three waves of introduction of new gauges reflected developments in engineering practice that originated in Britain and then spread to North America. The first wave reflected the uncritical transfer of practice from primitive mining railways to the earliest commercial railways. Subsequent waves were attempts to improve on prior practice.

The Early Gauges

A variety of gauges were used for the primitive railways that developed in the mining districts of Britain during the late eighteenth century, with 4'8" (1422 mm.) prevalent in the region near Newcastle where engineer George Stephenson performed early experiments with steam locomotion and built the coal-carrying Stockton and Darlington Railway. Stephenson was chosen to build the Liverpool and Manchester (L&M) Railway between 1826 and 1830. As the first railway designed exclusively for steam locomotion and the first substantial railway to rely exclusively on commercial and passenger rather than mining traffic, the L&M embodied significant advances in railway technology. Nevertheless, Stephenson used the same gauge as before, except for adding half an inch (13 mm.) between the rails to allow for more space between rails and wheel flanges.[11] By contrast, Stephenson's main rivals for the contract to build the L&M planned to use an unprecedentedly broad gauge, 5'6", as a reflection of what they regarded as a new engineering problem.[12]

Many North American engineers visited the L&M and other early British railways, regarding the L&M in particular as representing the best practice of its day. These engineers and their protégés were responsible for a large majority of the U.S. route mileage constructed during the 1830s. As a result, they "transferred virtually an entire technological 'package' to the United States," including the gauge.[13] Three engineers from North America's first commercial railway, the Baltimore and Ohio (B&O) Railroad, visited England in 1829 and increased their planned gauge from 4'6" to 4'8.5" in order to "admit cars of the dimensions used upon the Liverpool and Manchester rail way," to improve stability and equalize "the bearing on the wheels," and to provide more space for the locomotive apparatus.[14]

Unfortunately for later network integration, some U.S. engineers drew from the L&M a lesson only about the approximate gauge technically suitable for a railway. Thus, an 1832 report of New York state's railroad commissioners—citing the B&O's example—said that "the distance between the two tracks, for the wheels, should be about five feet."[15] As a result of this way of thinking, four gauges arose within a range of 3.5", all introduced by engineers who had visited England. The 5'0" gauge was chosen in part because it was a round measure. Given the wide engineering tolerances of early railways, the 4'9" gauge was practically interchangeable with 4'8.5", but exchange of traffic with 4'10" was difficult, if barely possible. Exchange between the narrower gauges and 5'0" was impossible.

Broad Gauges

Beginning in the mid-1830s, some British locomotive builders found their ability to develop powerful, easily maintained engines constrained by the 4'8.5" gauge, while certain civil engineers expected a broader gauge to promote improved stability, smoothness of ride, speed, and capacity. As a result, Britain's extensive Great Western Railway system adopted a gauge of fully 7'0" (2134 mm.), and a few short lines adopted 5'0" and 5'6".

North American engineers learned of British developments through both personal visits and the trade press, and their own experience also encouraged experimentation with broad gauges. The 6’0” gauge of the New York and Erie (hereafter, Erie) Railroad, whose route traversed the southern tier of counties in New York state from the Hudson River to Lake Erie, was proposed in 1836 by surveying engineer Edwin F. Johnson as a measure that “would permit the greatest speed and accommodation, at the least cost.”[16] In 1838, new chief engineer H.C. Seymour endorsed the gauge in deference to the expertise of English broad-gauge advocates,[17] and the first section of track opened at that gauge in 1841.

Later broad-gauge proponents supported more moderate gauges. Two veterans of the Erie Railroad introduced the gauge of 5’6” both to Canada (with Maine) in 1848 and Missouri in 1852. In the first case, a route connecting Canada’s commercial center, Montreal, with the ice-free port of Portland, Maine, chief engineer A.C. Morton prepared a report that summed up the advantages of a broader gauge as “an enlarged capacity” for both passengers and freight, leading to better service and reduced costs, and he reported that nearly all the U.S. locomotive builders he had surveyed desired a gauge of 5’0” to 5’6”.[18] Canadian promoters of the route accepted Morton’s proposal as reflecting current best practice in Britain.

Narrow Gauges

During the 1840s and 1850s, new locomotive designs reduced the technical advantage of broad gauges for power, and in the 1860s it became practical to build locomotives for gauges substantially narrower than 4’8.5”. As a result, engineers in Norway, Australia, and India experimented with a 3’6” gauge. Narrow gauges have a sometimes useful advantage in facilitating sharper curvatures, enabling railway lines to follow the contours of rugged landscape and reducing the need for expensive cuttings, tunnels, embankments, and bridges. Around 1869, some British engineers began to argue, much more dubiously, that railway construction and operating costs generally are proportional to gauge.

Railway promoters in Toronto adopted the British colonial gauge of 3’6” for two railways in lightly populated districts of Ontario. Narrow-gauge advocates in the United States were influenced by British engineers who supported still narrower gauges. Engineer and promoter William J. Palmer became convinced that his Denver and Rio Grande Railway in Colorado could be built at a 3’0” gauge for one third less initial cost than at 4’8.5” and have lower operating costs as well. The first section of this line, like the Canadian lines, opened in 1871.

Incentives Governing the Diffusion of Gauges

Track gauges in North America were chosen by the individual private companies, with government regulation affecting only a few cases. Preferences for specific gauges affected choices primarily in regions without previous railways, while an interest in gauge compatibility dominated choices where neighboring lines already existed.[19] In later years, expectations played an increasing role.

Preferences in New Regions

Promoters of lines in new regions generally deferred in their choice of gauge to their chief engineers, and the engineers generally held to the gauge they had used in previous construction. Thus the engineers who built the B&O went on to use 4’8.5” for the first railways in New England. Similarly, the early 4’8.5” railways of Virginia, Louisiana, Kentucky, and later Texas were built by engineers known to have used the gauge before.[20]

The gauge of 4’8.5” was the one most often adopted in new regions, especially during the 1830s (table 1). The 4’10” gauge of New Jersey was transplanted early to the states of Ohio and Mississippi, and the 5’0” South Carolina gauge both to nearby Florida and, later, to the first two railways on the west coast. The introduction of 4’10” to Ohio was the result of buying a New Jersey locomotive “off the shelf.” Promoters of Ohio’s Mad River and Lake Erie Railroad were reportedly impressed with the novel whistle of a locomotive available as a result of a canceled order. They bought it and built their track to fit.[21]

[Table 1 here]

The Erie Railroad’s 6’0” gauge was adopted in 1853 by the first 32-mile line leading westward out of Chicago due to a relocated engineer—the original proposer of the gauge, Edwin F. Johnson.[22] As already noted, two other veterans of the Erie also introduced broad gauges to new regions, although they used the variant measure of 5’6”.

The diffusion of narrow-gauge practice involved less the influence of itinerant engineers than the rapid adoption by promoters of a new engineering “doctrine” based on unrealistic expectations of cost savings. Opponents likened this movement to a "fever." The example of Palmer's 3'0" line and other early routes played a role, as did a large pamphlet literature. Hundreds of lines arose in the empty spaces between established broader-gauge routes.[23]

The Interest in Gauge Compatibility

Railways' interest in compatibility of gauge was weakest at the very beginning of North American railways but grew over time. Diversity was initially not perceived to be costly, because few expected that the railways would link up in an interconnected, traffic-exchanging network. The earliest commercial railways served only a local, or at best regional, purpose: to connect cities with their hinterlands or the western waterways, or to link water routes over territory that was unsuitable for canals.

Nevertheless, even in the earliest years new railways nearly always adopted the same gauge as an established connecting or adjacent line, if one existed, or even a line that entered the same city or that lay across a river which could be bridged or ferried. Part of the reason for this must often have been the common influence of a particular engineering tradition, but the rarity of exceptions points also to an apparent desire to make exchange of traffic possible.

The few exceptions to this rule before 1870 are noted in table 1. Two of these involve lines that entered Great Lake port cities from different directions than established lines and thus presumably expected no exchange of traffic with them. Four other exceptions involve lines that had reason to expect future connections with other lines of their own chosen gauge.

This leaves two exceptional cases in which the promoters of new railway lines believed that the normal presumption favoring the use of a common gauge did not apply to them. They were building, so they thought, self-contained railway systems. The New York City promoters of the Erie Railroad sought to monopolize the railway traffic of southern New York state. Erie promoters had secured a state charter in 1832 that "expressly prohibited, under penalty of its forfeiture, connection with any railroad leading into New Jersey, Pennsylvania, or Ohio, on the childish theory that thus traffic could not be diverted from the Erie," but when their engineers proposed the adoption of a variant gauge, they saw gauge diversity as a better assurance against this diversion.[24] The promoters believed initially that their western terminal on Lake Erie would suffice to compete for traffic from the Midwest.

Similarly, the Montreal interests who promoted a route to Portland were interested primarily in an ice-free port for trade with Britain, and they sought to keep Canadian commerce independent of Boston or other leading U.S. commercial centers. Thus they had little positive interest in common-gauge connections with them. The Portland promoters of the route had a strong interest against such connections. Eager to prevent diversion of Canadian traffic to Boston, they had secured a state charter prohibiting future connections on the western side of the line, and they welcomed the proposal of a variant gauge.[25] Taylor and Neu go so far as to argue that engineer Morton's claims of technical superiority for the broad gauge "were very largely window dressing" for a decision to create deliberate breaks of gauge between Montreal and Boston.[26] Taylor and Neu appear unaware of the extent to which engineers truly supported broad gauges. While Portland promoters of the route showed an interest in these breaks of gauge as such, the Montreal promoters did not.

An interest in gauge compatibility is seen a few years later in the case of Canada's Great Western Railway (GWR), a route across southwestern Ontario from the New York border at Niagara to the Michigan border at Windsor-Detroit. Allied with—and partly financed by—U.S. interests seeking a common-gauge link between the northeastern and midwestern states, the GWR planned to adopt 4'8.5". However, Canadian legislators had learned from events in Britain the importance of a standard gauge, and they treated the GWR as part of the future Canadian network that would expand from Montreal, not as part of a U.S. network. They imposed 5'6" as a standard in 1851, and the GWR opened at that gauge in 1854.

The gauge of the first U.S. transcontinental railroad was also fixed by law in a way that reflected cooperation to produce gauge compatibility. Congress had originally left choice of gauge to President Lincoln, who decided on 5'0" on the advice of an engineer with broad-gauge experience.[27] Midwestern railways and their Eastern allies then successfully lobbied Congress to specify 4'8.5" as the system's gauge.[28]

Another strong indication of a growing interest in compatibility was the choice of 5'6" by widely separated lines in Missouri, Texas, Louisiana, and Arkansas (table 1). The Pacific Railroad of Missouri selected its gauge in 1851 not only on the basis of technical considerations but also due to the expectation that the lower Mississippi and Missouri Rivers were unbridgeable, so that compatibility with established eastern railways was less relevant.[29] Adoption of the same gauge in the other states suggests a self-fulfilling expectation of the emergence of a separate regional standard.[30] The 5’6” system taking shape in Missouri was the most substantial network that the new lines might expect to connect to, and St. Louis was certainly the most prized destination. After two lines in Texas adopted the gauge, three railways in western Louisiana building toward Texas did the same. A few years later, Arkansas’ pioneer railway followed the lead of these lines in neighboring states.

In the building of narrow-gauge railways after 1870, breaks of gauge with established lines became more the rule than the exception. William Palmer, the narrow-gauge pioneer in Colorado, sought to develop an integrated regional railway system with few points of contact with broader-gauge lines. Edward Hulbert, a prominent southern railway executive, vigorously advocated narrow-gauge branch lines as a means of economic development for areas where standard-gauge lines were uneconomical. Both thus accepted transshipment costs as a price to be paid for expected greater savings in transport costs. Some narrow gauge advocates, however, went so far as to pursue an entire alternative nationwide system of trunk lines that would eliminate the need to connect with broader-gauge railways.[31]

Whether or not individual promoters expected the development of a national network, they at least had learned the value of reducing diversity among connections that do develop. Thus, one result of the first National [U.S.] Narrow-Gauge Railway Convention, called in 1872 by Hulbert, was a resolution to adopt 3’0” as the standard narrow gauge.[32] More than 95 percent of U.S. narrow-gauge lines in fact then used that gauge.[33]

The Emergence of Common-Gauge Regions

The overall prevalence of 4’8.5” that emerged even prior to large-scale conversion is not surprising given the number of local origins of that gauge, but the emergence of substantial regions of other gauges can be explained only by looking at the specific path of events.

The 4’8.5” Networks and 4’10” in New Jersey

Each of the three major 4’8.5”-gauge regions had multiple local origins in smaller networks that eventually merged into one another. The mid-Atlantic region began with the Baltimore and Ohio Railroad and also with railways in eastern Pennsylvania, southern Virginia, and eastern North Carolina. In the northeast, 4'8.5" networks spread out from Albany, from Boston, and from New York City. The Midwestern region began with routes at the border of Ohio with Michigan and in southern Indiana. The 4'10" network in New Jersey expanded from the first line built at that gauge, the Camden and Amboy.

5'0" and other Gauges in the South

The 4'8.5" gauge was introduced by a substantial majority of the local pioneering lines in the South (table 1), and Mississippi's first railway adopted 4'10". Nevertheless, it was the 5'0"-gauge Charleston and Hamburg (C&H) Railroad that set the standard for the South as a whole. The reason for this was that none of the narrower lines expanded into substantial networks, while the merchants of Charleston used a series of regional railroad conventions to promote a route from Charleston to the Mississippi.

The separately managed lines in this route adopted the 5'0" gauge as they built westward through Atlanta and Chattanooga, reaching the Mississippi at Memphis in 1857. The route became the backbone of a Southern network, as railway promoters seeking long-distance connections tapped into it with routes of the same gauge from Savannah during the 1840s; from southern Virginia during the early 1850s; and subsequently from Mobile, New Orleans, Nashville, and Louisville. As a result, the 5'0"-gauge network was built on top of the earlier 4'8.5" network of southern Virginia and the earlier individual 4'8.5" and 4'10" lines elsewhere in the South. As the 5'0" gauge spread, construction at other gauges ceased. Railway builders in North Carolina were prohibited by law from using 5'0", as the state sought to use gauge differences to channel the state's commerce to its own coastal ports.

Taylor and Neu found it "logically incomprehensible" that Southern lines outside the border states adopted gauges other than 5'0".[34] A simple knowledge of the order of events and dynamics of the process, however, makes it very comprehensible indeed.

4'10" and Other Gauges in Ohio

Although the first two railways in Ohio adopted 4'8.5" and 5'4", a large majority of the state's antebellum railways adopted the 4'10" gauge of the third line built, the Mad River and Lake Erie (MR&LE) Railroad. As in the case of the South, the explanation for this rests on the order of subsequent construction.[35] Even before the MR&LE was completed in 1848, a connecting line was built to the Ohio River at Cincinnati. Access to Cincinnati stimulated extension of the network eastward to Columbus in 1850 and to the northeastern commercial center of Cleveland in 1851. Further connecting lines spread the 4'10"-gauge region through most of the rest of Ohio and part of eastern Indiana during the early 1850s and along Lake Erie to Buffalo, New York by 1854.[36] Pittsburgh and Chicago were reached by 1858, in the latter case by means of lines built across a mostly 4'8.5"-gauge region.

Ohio's southern rim east of Cincinnati was unreached by the 4'10" network, and here the Marietta and Cincinnati (M&C) Railroad adopted 4'8.5" because of a partnership with the Baltimore and Ohio Railroad, which opened a new western branch to the Ohio River opposite the M&C's eastern terminus. Both lines opened in 1857.

The Broad-Gauge Networks

Branches of the New York and Erie Railroad soon extended through much of southern New York state and the coal fields of northeastern Pennsylvania. By 1853, the Erie's managers recognized that the company also needed direct rail connections, and not merely a Lake Erie terminus, to compete effectively for the commerce of the Midwest. The Erie therefore pursued alliances with several proposed railways that together formed two routes to the Mississippi, one ending opposite St. Louis and the other passing through Chicago.[37] The lines expected to benefit not only from the technical advantages of the 6'0" gauge but also from being able to offer the only common-gauge routes from the seaboard to the Midwest.

However, all of these allies reneged on the arrangement except the Ohio and Mississippi Railroad, opened from Cincinnati to East St. Louis, Illinois in 1857. As a result, the Erie itself organized (and partly owned) a connecting route which opened in 1864. This route comprised both a new line to Dayton, Ohio and running rights from there to Cincinnati over a 4'10" railway, newly straddled by broad-gauge rails. The Erie system was the strongest example in its time of explicit cooperation among railway companies to pursue a common-gauge route.

Meanwhile, the Montreal railway promoters extended their 5’6”-gauge system, renamed the Grand Trunk Railway, westward through Toronto to the Michigan state border south of Lake Huron by 1858. Numerous branch lines filled out an extensive network.

The 5’6”-gauge lines in the southwest never came together in a single network.

Narrow-Gauge Networks

The Denver and Rio Grande Railway and allied lines branched out into much of central and southwestern Colorado, northern New Mexico, and eastern Utah, reaching Salt Lake City by 1883. The network included 2,783 route miles at its peak in the mid-1880s.[38] Elsewhere, the largest narrow-gauge networks of the 1870s and 1880s emerged in the oil fields of northwestern Pennsylvania, in central Florida, and on Canada’s Prince Edward Island. Later, a 3’6” network in Newfoundland reached a peak length of 969 miles in the 1910s.

The financial failure in 1883 of two companies attempting to form a “Grand Narrow Gauge Trunk” from Ohio to Texas was interpreted as demonstrating the infeasibility of competing with broader-gauge routes. This, combined with growing disillusionment over the supposed cost savings of the gauge, brought about a collapse of the “narrow-gauge fever.”[39]

The Resolution of Gauge Diversity

In the 20 years from 1866 to 1886, the various regional standard gauges gave way to a nearly continent-wide standard, 4’8.5”. Why did standardization happen at this time? Did this gauge “win” because it was the tested and proven superior gauge? Or was its selection the continued outworking of the history of gauge choices?

Engineers of the time discussed the relative merits of 4’8.5” and the broader gauges only rarely. Some favored 4’8.5”, some a broad gauge, others found essentially no difference, but none suggested that the discussion mattered for the question of which gauge should become the continental standard.[40]

Standardization was stimulated by an increase in traffic demand, and the pattern of this increase favored the gauge of 4’8.5”. During the 1850s, much of the relatively high-valued traffic from the industrial East to the agricultural Midwest shifted from water routes to rail, but the low-valued bulk grain traffic from west to east still used rail only for peak loads or in winter. By 1870, about half this voluminous and growing traffic went by rail. Breaks of gauge thus became increasingly costly, and network integration more valuable.

Most points of origin and most destinations for these shipments used the gauge of 4’8.5”. In between, however, shipments crossed either the 4’10”-gauge region centered on Ohio or the 5’6” region of southwestern Ontario. Thus those regions had the highest payoffs to conversion. When they did convert, the resulting common-gauge region held a substantial majority of North American railway mileage and traffic, making it the core of a developing continental network.

Growing demand for interregional traffic also spurred the development both of interregional systems under common management and of cooperation among railway lines.[41] These systems and cooperating groups facilitated standardization, because they internalized the externalities of the gauge choices of individual local lines. Such mechanisms as side payments, appropriation, and simple coordination thus led to gauge conversions that would have been delayed if each local line had acted simply on the basis of its own incentives. As Liebowitz and Margolis would predict, these mechanisms thus reduced the inefficiency resulting from earlier lack of foresight.

The costs of conversion, it should be noted, were not high in relation to the value of railways. Most of the cost of original construction went into the preparation of the roadbed, and this did not need to be alterred except in the widening and upgrading of lightly built narrow-gauge track. In cases where rails were spiked directly to wooden cross-ties, conversion of track was often a relatively simple matter of pulling up spikes, shoving one rail to a new position, and setting new spikes. Some parts of switch assemblies and crossings required replacement, and in some locations it was necessary to move both rails. Some broad-gauge lines moved both rails in order to keep weight centered on the roadbed.

In one of the best documented cases, it cost an estimated $1,066 per mile to reduce the track of the 6’0”-gauge Ohio and Mississippi Railroad to 4’8.5” in 1871, plus an additional $5,060 per locomotive and from $518 to $1,380 apiece for other rolling stock.[42] At the low end of costs, conversion of the relatively lightly built southern railways in 1885 and 1886 cost an estimated $150 per mile for both track and equipment.[43] Conversion of one narrow-gauge line "with minimal improvements" cost $7,500 per mile.[44]

Railways also reduced the cost of diversity by using adapter or "gateway" techniques—practices that permitted a degree of network integration even when gauges differed.[45] Mixed-gauge (three- or four-rail) track was used for gauges differing by at least several inches, while "compromise gauge" wheels or track were used in the narrow range between 4'8.5" and 4'10". Numerous railways exchanged the bogies (wheel sets) of rolling stock using either hoists or specially designed pits, and one major line experimented with adjustable wheels.

Intra-Regional Standardization

The resolution of gauge diversity began on a small scale in the decades before the U.S. Civil War with the conversions of 13 local lines totaling 320 miles. In each case, the line had unexpectedly gained the opportunity to exchange traffic with a substantial network of a different gauge. The pace of remedying intra-regional diversity picked up during the late 1860s with the conversion of several variant-gauge Southern lines to 5'0", followed by the conversion of three long routes in North Carolina in 1879.

The 5'6" Network of Canada

Of greater importance was the resolution of diversity among the major common-gauge regions. As the first step in this process, Canada's Great Western Railway laid a 4'8.5" third rail during 1865 and 1866 on its route between New York and Michigan, offering the first connection among the separated regions using that gauge. The New York Central and Michigan Central Railroads made a side payment for a substantial share of the cost.[46]

The GWR's action put the Grand Trunk Railway at a disadvantage in competing for traffic on its own routes between the northeastern and midwestern United States. In response, the Grand Trunk experimented, beginning in 1869, with adjustable-gauge wheels for cars used in international traffic. However, the cars did not function well, and it became clear that this substitute for standardization would not be a lasting solution to the diversity of gauge.

In 1872, Grand Trunk management concluded that the broad gauge had yielded none of its expected technical advantages, but rather additional expense due to breaks of gauge.[47] Because new Canadian railway companies were proposing to use 4'8.5" for routes both to the

Pacific coast and to the maritime provinces, the Grand Trunk decided to convert. The company's direct route between New York and Michigan was changed in November 1872, the line east to Toronto, Montreal, and Portland about a year later. Several hundred additional miles of other companies' lines converted gradually by 1885, and one final short line in 1911.

The 4'10" Networks of Ohio and New Jersey

As the first step in integrating the 4'10"-gauge network of Ohio with the 4'8.5"-gauge networks both east and west, Ohio railways began allowing passage of narrower-gauge rolling stock during the late 1850s. This practice proved workable using wheels with broader than normal treads, but oscillation of the cars damaged the track. As cross-gauge traffic grew during the late 1860s, the Ohio lines began insisting that the cars use a slightly wider compromise gauge. This, however, increased friction and costs on 4'8.5"-gauge track.[48]

In 1869, the Pennsylvania Railroad (PRR) took long-term leases of three 4'10" lines offering routes across Ohio to Chicago and other western destinations. In order to integrate these routes into a system, the PRR increased the gauge of its original route east of Pittsburgh by half an inch to 4'9" and reduced the gauge of the western routes to 4'9.5". This reduced a problematic 1.5" gauge differential to a series of manageable half-inch differences.

Another externality-internalizing system formed in 1869, the Lake Shore and Michigan Southern (LS&MS) Railroad, converted both its 4'10" route from Buffalo to Cleveland and its 4'8.5"-gauge route from Cleveland to Toledo to 4'9.5". The independent Ohio railways followed the lead of the emerging systems, adopting 4'9.5" or other intermediate gauges during the early 1870s. The change was encouraged by limits that the trunk-line systems imposed on the wheel gauges of rolling stock accepted for through shipments. Between 1877 and 1879, the cross-Ohio trunk routes reduced their gauges again to either 4'9" (PRR) or 4'8.5" (LS&MS), and by 1880 only two short lines retained 4'10".

The expanding Pennsylvania Railroad system also took the lead in narrowing the gauge in New Jersey. In 1871 the PRR took a 999-year lease of several routes connecting Pennsylvania with the New York City area. Conversion to 4'9" followed quickly. Most of the independent 4'10"-gauge lines in the region converted to either 4'8.5" or 4'9" by the end of

1874.

Interestingly, as the PRR system became probably the most important set of trunk routes in the continent during the 1870s and 1880s, its variant 4’9” gauge was adopted by other railways far beyond the system’s own boundaries. The American Railway Association discussed the resulting diversity in 1896 and 1897 and, by non-binding resolution, affirmed 4’8.5” as the standard gauge. Most wider track was adjusted in the following several years.[49]

The 5’6” Network of the Southwest

Conversion of the 5’6”-gauge lines in Missouri and states to its south followed the ending of the region’s physical isolation, as the Missouri River was bridged at Kansas City in 1869 and a bridge across the Mississippi at St. Louis was then well underway. The Pacific Railroad of Missouri, which joined those two cities, converted its 318-mile route to 4’8.5” in July 1869. At about the same time, both the St. Louis and Iron Mountain Railroad in Missouri and the Memphis and Little Rock Railroad in Arkansas established car-ferry service across the Mississippi with the Southern-gauge network. As these lines lacked direct connections with lines of other gauges, they converted to 5’0”. However, both lines converted again in the late 1870s to 4’8.5” after neighboring lines west of the Mississippi were built at that gauge. Similarly, the broad-gauge lines in Texas converted to 4’8.5” in 1876 and 1877 when the expanding network of that gauge reached the state.

The 6’0” Network

Conversion of the 6’0”-gauge system of the Erie Railroad and its branches began on the periphery before proceeding to the core. In 1866 and 1871 two 4’8.5”-gauge systems in northern Pennsylvania expanded into New York state by acquiring 6’0”-gauge branch lines, converting the gauge of one and laying a third rail on the other. These actions established the first direct links between the mid-Atlantic and northeastern regions of the 4’8.5” gauge.

In 1871 the Ohio and Mississippi Railroad, linking Cincinnati to East St. Louis, converted to 4’8.5”. The broad-gauge route from New York to the Mississippi had never functioned in an integrated fashion, and the O&M had come to exchange much traffic with the Baltimore and Ohio Railroad, by way of the Cincinnati and Marietta. When the B&O bridged the Ohio

River in 1871, and with the bridge across the Mississippi into St. Louis under construction, the O&M's opportunity to gain long-distance 4'8.5" connections was decisive.

In 1873, the Erie Railroad's management concluded that the line could not compete effectively for through traffic without using 4'8.5", and the company began laying a third rail along the entire trunk route in New York state. As a shortage of funds during the recession of the mid-1870s delayed completion of this project until 1878, the Erie made extensive use of hoisting machines to exchange bogies in the meantime. When completed, the Erie trunk's third rail enabled its branch lines—comprising over 20 companies and 1,500 miles—to convert to 4'8.5" at their own pace. By maintaining some broad-gauge service until the early 1880s, numerous unconvertible locomotives and other equipment could continue in service until the end of their useful life, reducing the ultimate cost of conversion.

The 5'0" Southern Network

Although the South's internal network integration improved greatly after the Civil War, it was left out of an increasingly integrated, common-gauge network spanning the rest of North America. Rail traffic with other regions grew substantially during the 1870s and 1880s. The cost of breaks of gauge grew in proportion.

The South's conversion began as a "bandwagon" process, in which lines that valued conversion more highly switched first, thereby increasing the value of conversion for other lines and inducing them to follow.[50] On the northern edge of the South, the Kentucky Central (KC) Railroad converted in July 1881 when completion of the 4'8.5" Chesapeake and Ohio (C&O) Railroad made the KC a link in carrying Midwestern grain to the Virginia seaboard. That same month, the interregional Illinois Central system improved its internal integration by converting its 548-mile southern route from the Ohio River to New Orleans. This action induced the conversion in 1882 of the 392-mile Chesapeake, Ohio, and Southwestern Railroad, which linked the C&O with the Illinois Central and also met the 4'9" gauge at Louisville. In 1885, the 493-mile Mobile and Ohio Railroad, a longtime partner in traffic with the Illinois Central, converted as well.

These conversions created numerous breaks of gauge within the western South and put competing interregional routes at a disadvantage in bidding for traffic. Meanwhile, the long-standing breaks of gauge between 4’8.5” and 5’0” lines in Virginia and North Carolina impeded the growing volume of traffic along the seaboard.

It became apparent in 1885 that several additional Southern railways were considering conversion. In response, managers of all the major Southern lines held informal discussions and decided to convert simultaneously, thus maintaining their intraregional integration while improving links beyond the region. A formal conference in February 1886 made concrete plans, choosing the Pennsylvania Railroad’s 4’9” gauge and setting May 31 and June 1 as the main dates for track work. Over 13,000 miles of track were converted on those dates, and over 1,500 miles of branch lines and sidings within a few days before and after. The rolling stock was converted over a period of several months before and after the track work.[51]

Narrow-Gauge Networks

After U.S. narrow-gauge mileage reached a peak of 11,669 in 1885, over 2,300 miles changed gauge during 1887 and 1888 alone, but further conversions and abandonments came slowly. On the extensive Denver and Rio Grande system, competitive threats from new 4’8.5”-gauge railways induced conversion of the entire trunk line from Denver to Salt Lake City between 1888 and 1890. An extensive network of narrow-gauge D&RG routes and independent lines, centered in southwestern Colorado, lasted well into the twentieth century. By 1969, all but 46 miles went out of service. At present, less than 100 miles of U.S. narrow-gauge railways remain in service, all as tourist lines. Newfoundland’s 3’6”-gauge system remains in use for general traffic.

Discussion

Four feet 8.5 inches became the North America standard gauge because its multiple early adoptions by lines in the northeastern, mid-Atlantic, and midwestern United States made it the standard gauge of those regions. When demand grew for through-routes between the seaboard and the Midwest, the gauge displaced the original gauges in the intervening regions of Ohio and southwestern Ontario. The now unified 4’8.5”-gauge region formed the core of an expanding, integrated continental network, and the remaining common-gauge regions converted in order to participate in the advantages of network integration.

Ultimately, 4’8.5” became the standard gauge because most of the earliest railway builders accepted British example as embodying best practice in railway engineering. Subsequently, a desire to maintain compatibility with the gauge’s large installed base continued to make it the most common choice of new lines.

Even in the earliest years, chance preferences for the slightly varying gauges of 5’0” and 4’10” by lines in South Carolina and Ohio led to the adoption of those gauges over substantial regions. In each case, it happened to be that line, rather than another using 4’8.5”, that became the nucleus of an expanding regional network. If either line had adopted 4’8.5” from the start, North American railways would have had much less diversity to resolve. On the other hand, if a few key early 4’8.5” lines had chosen a different gauge instead, then there could have been much more diversity—and conceivably a different continental standard.

In later years, gauge variation became a deliberate strategy, as railway builders introduced supposedly superior gauges, both broad and narrow, for supposedly self-contained regions. These strategies failed, both because the gauges proved to be less advantageous than supposed and because larger-scale network integration proved more important than expected.

As demand for interregional transport grew, market incentives led private-sector agents to resolve early diversity into a continental standard. Private ownership facilitated this through side payments and formation of externality-internalizing interregional systems.[52] Considering that the actual diversity was resolved at a relatively early point both in the growth of traffic demand and in the building of systems, one may reasonably believe that even a much greater diversity—had it developed—could have been resolved by the later, stronger incentives.

This fact suggests two possible reasons why diversity has still not been resolved in other parts of the world.[53] First, interregional traffic demand may be low relative to the (in some cases immense) potential cost of conversion. Second, institutional factors may prevent the internalization of externalities, perhaps due to ownership of railways by different states.

There is no reason to see 4’8.5” as technically or economically optimal for railways, whether given the operating conditions of 1830, 1890, or 2000. The gauge was little tested before its initial adoption, and its continued use has been the result of its prior use. Still, alternative gauges did not prove greatly superior historically. And although many engineers today regard broader gauges as better, they do not see potential gains as large.

What, then, are the lessons of the case of North American track gauge for the theory of path dependence? First, the case of track gauge illustrates one way in which an allocation process may be path dependent—not a predictable convergence to a uniquely optimal technique but rather a historically contingent selection among a range of possibilities. Agents acted purposefully, optimizing on the basis of their limited understanding of the relative merits of different gauges and their at first limited foresight into the value of standardization and network integration. "Historical accident" operated as well, particularly insofar as the initial choices of gauge in new regions, and the growth of common-gauge networks outward from one early line rather than another, were based on idiosyncratic persons and events. In line with Arthur's models, early events generated positive feedbacks that affected later events. In line with Liebowitz' and Margolis' analysis, these early events depended in part on lack of foresight, and later purposeful, profit-seeking actions substantially altered the initial outcome. Still, today's standard gauge is the result not only of rational purpose, but also of nonpurposeful events, of the unforeseen consequences of purposeful action, and of the underlying dynamic structure of the allocation process.

On the issue of efficiency, the case of track gauge illustrates Arthur's contention that a path-dependent process, in contrast to standard neoclassical processes, does not necessarily select, or converge to, the theoretically most efficient outcome. The gauge selection process was "path inefficient," in that alternative potential paths of allocation would have generated higher payoffs—that is, lower costs of enduring and resolving diversity. The case of gauge also adds a facet to Arthur's concept of inefficiency. In Arthur's non-spatial models, relative inefficiency arises when the selected technique has lower payoffs than other available techniques would have yielded. Here, in a spatial process, inefficiency arose primarily because standardization took place initially at only a local rather than continental level.

The case of gauge also offers support for Liebowitz and Margolis' proposition that, as inefficiency is revealed, it can generate profit opportunities that draw forth behavior to remedy it. Liebowitz and Margolis point particularly to coordination and appropriation, mechanisms that contributed historically to the resolution of gauge diversity. The case of gauge also adds a facet to Liebowitz' and Margolis' discussion of appropriation, in that here it was agents themselves that were appropriated, not the competing techniques.

There is no contradiction between these views of Arthur and of Liebowitz and Margolis on efficiency. Arthur measures the results of an allocation process against a theoretical ideal; Liebowitz' and Margolis' standard is what can be achieved by purposeful behavior. Perhaps the most important unresolved issue is how great the inefficiency of a path-dependent process can be. The inefficiency in North American track gauge was by some standards rather small, both because the efficiency of railways was not highly sensitive to gauge and because conversion costs were small.[54] For other technologies, the choice of a specific technique may be much more crucial and switching costs greater.

Another unresolved—and disputed—issue is whether path-dependent processes can be a source of market failure.[55] The inefficiency in track gauge selection was not primarily a matter of market failure—it was not driven chiefly by a difference between private and social costs and benefits.[56] Rather, it was driven by imperfect foresight into both the future value of interregional gauge standardization and the relative performance of different gauges.

Were the costs of gauge diversity then simply learning costs? No. Learning costs are properly the fixed, sunk costs of discovering the relative advantages of different techniques. The costs of enduring and resolving gauge diversity—of having started on a suboptimal path of allocation—continued even after the learning had taken place. These costs were much more the costs of initial ignorance than of subsequent learning. This point is perhaps clearer in countries and continents where costly diversity never was resolved. Railways in these cases learned the benefits of standardization, but not in a way that they have been able to use.

In closing, the case of North American track gauge also shows that foresight itself is a factor that varies among agents. One journalist in 1832 showed unusual foresight both into the consequences of introducing a variety of gauges for different local projects and also into the mechanism that would drive the dynamics of future gauge selection. According to an unsigned commentary in the American Rail-road Journal, "when we consider that most of the principal Rail-ways now in progress ... must soon intersect each other, either by the extension of present lines or the formation of new ones, we are forced to conclude that this discrepancy in the width of tracks, will ultimately produce an infinitude of vexations, transfers and delays which might easily have been avoided," as each company could have adopted "the mean width of the whole without any possible detriment." The author argued that "the establishment of a particular width, by statute, in two or three of the principal States" or as a convention by a few major railways "would probably have influence sufficient to produce the desired uniformity in most cases throughout the United States."[57] Unfortunately, neither the railways nor their potential regulators possessed this author's foresight.

REFERENCES

American Rail-road Journal 1 (January 21, 1832).

Arthur, W. Brian. “Competing Technologies, Increasing Returns, and Lock-in by Historical Events.” Economic Journal 99 (1989): 116-31.

—. Increasing Returns and Path Dependence in the Economy. Ann Arbor, 1994.

Baer, Christopher T., ed. Canals and Railroads of the Mid-Atlantic States, 1800-1860. Wilmington, Delaware, 1981.

Baltimore and Ohio Railroad Co. Annual Report 4 (1829-30). Baltimore.

Carlson, Robert E., The Liverpool and Manchester Railway Project 1821-1831. New York, 1969.

Currie, Archibald W. The Grand Trunk Railway of Canada. Toronto, 1957.

David, Paul A. “Clio and the Economics of QWERTY,” American Economic Review 75, no. 2 (1985): 332-37.

—. “Path-dependence and Predictability in Dynamic Systems with Local Network Externalities: A Paradigm for Historical Economics.” In Technology and the Wealth of Nations: The Dynamics of Constructed Advantage, edited by D. Foray and C. Freeman. London: Pinter, 1993.

Farrell, Joseph and Garth Saloner. “Standardization, Compatibility, and Innovation,” Rand Journal of Economics 16 (1985): 70-83.

Great Britain. Parliament. Report of the Gauge Commissioners. London, 1846.

Haney, Lewis H. “A Congressional History of Railways in the United States.” In Economic and Political Science Series 3 (1906). University of Wisconsin.

Hilton, George W. American Narrow-Gauge Railroads. Stanford, 1990.

Johnson, Edwin F. Transcontinental Railways of the United States. Hartford, 1871.

Kirkwood, James P. Report on the Guage [sic] of Track to the Board of Directors of the Pacific Railroad, Missouri. St. Louis, 1854.

Lavallée, Omer. Narrow Gauge Railways of Canada. Montreal, 1972.

Lovett, Henry A. Canada and the Grand Trunk, 1829-1929, reprint edition. New York and Toronto, 1981.

Liebowitz, S.J., and Stephen E. Margolis. "The Fable of the Keys." Journal of Law and Economics 33 (1990): 1-25.

— and —. "Network Externality: An Uncommon Tragedy." Journal of Economic Perspectives 8, no. 2 (1994): 133-50.

— and —. "Path Dependence, Lock-In, and History." Journal of Law, Economics, and Organization (1995).

Mobile and Ohio Railroad Co. Annual Report 37 (1984-85). New York.

Morton, A.C. Report on the Gauge for the St. Lawrence and Atlantic Rail-Road. Montreal, 1847.

Mott, Edward H. Between the Ocean and the Lakes: The Story of the Erie. New York, 1899.

National Narrow-Gauge Railway Convention. Proceedings of the National Narrow-Gauge Railway Convention, held at St. Louis, Missouri, June 19, 1872. St. Louis, 1872.

Pennsylvania Railroad Co. Annual Report 22 (1869). Philadelphia.

Poor, Henry V. Manual of the Railroads of the United States [title varies]. New York, annual beginning 1868/69.

Seymour, Silas. A Review of the Theory of Narrow Gauge as Applied to Main Trunk Lines of Railway. New York, 1871.

Shuman, Armin E. "Statistical Report of the Railroads in the United States." In Report on the Tenth Census, vol. 4: Report on the Agencies of Transportation in the United States. Washington, D.C.: U.S. Department of the Interior, Census Office, 1883.

Smiles, Samuel. The Life of George Stephenson and of his Son Robert Stephenson. New York, 1868.

Stapleton, Darwin H. "The Origin of American Railroad Technology." Railroad History 139 (1978): 65-77.

Taylor, George R., and Irene Neu. The American Railroad Network 1861-1890. Cambridge, Massachusetts, 1956.

[1]American Railroad Network, pp. 3-7.

[2]David, “Clio,” “Path-dependence;” Arthur, “Competing Technologies,” Increasing Returns; Liebowitz and Margolis, “Fable,” “Network Externality,” and “Path Dependence.”

[3]Source: interviews with engineers at the Association of American Railroads and the American Railroad Engineering Association. Hilton (American Narrow Gauge, p. 37) cites recent published views both that the optimal gauge is wider but “not far from” the current standard and that the optimal gauge is at least six feet.

[4] For example, according to David (“Clio”), “A path-dependent sequence of economic changes is one of which important influences upon the eventual outcome can be exerted by temporally remote events, including happenings dominated by chance elements rather than systematic forces.”

[5]These and subsequent data are compiled from a wide variety of sources. The most important of these are the 1880 U.S. transportation census (Shuman, “Statistical Report”), with gauges in that year and an annual construction series for 1,174 railroad companies; Baer, Canals and Railroads, with construction mileages, original gauges, and gauge conversions for the Mid-Atlantic states through 1860; the annual Poor’s Manuals, and similar yearbooks, which (like Baer) not only offer information on gauge but also enable the tracing of specific routes through company mergers and reorganizations; Canada Yearbook (1906), for annual construction in that country; and Hilton, American Narrow-Gauge Railroads and Lavallée, Narrow-Gauge Railways for data on U.S. and Canadian narrow-gauge railways, respectively. Construction dates in the U.S. census report, which sometimes reflect reorganizations or final completion of long routes, are corrected to the extent possible on the basis of other sources.

[6]The index value is the sum of squared shares of the mileages of the individual common-gauge regions. At each location, the proportion of continental route accessible without crossing a break of gauge is simply the share of that common-gauge region in continental mileage. Summation over all originating points is simply a matter of squaring and adding these proportions. Before 1850 the index cannot be meaningfully defined, as railways were too sparse to define larger common-gauge regions.

[7]In fact, network integration is a matter of the actual exchange of traffic, which a common gauge facilitates but does not assure. As Taylor and Neu discuss throughout their work, network integration was hindered until the early post-Civil-War years by unbridged (and unferried) rivers, lack of cross-town links between railways, and often an unwillingness to let rolling stock of one company pass over the track of another.

[8]Arthur, "Competing Technologies."

[9]Interested readers may contact me for a working paper.

[10]Liebowitz and Margolis, "Path Dependence." Arthur explicitly assumes conditions in his model under which foresight and some forms of coordination do not apply.

[11]Stephenson's son Robert later told a parliamentary commission that his father did not "propose" the gauge but rather simply "adopted" what was already in use in his home region (Great Britain, Report). Stephenson's friend and early biographer Samuel Smiles wrote that the gauge "was not fixed after any scientific theory, but adopted simply because its use had already been established" (Smiles, Life).

[12]Carlson, Liverpool and Manchester.

[13]Stapleton, "Origin."

[14]Baltimore and Ohio Railroad Co., Annual Report. The B&O experimented briefly during construction with a gauge of somewhat over 4'9", allowing additional space between the rails and wheel flanges.

[15]Quoted in Haney, "Congressional History," p. 198.

[16]Johnson, Transcontinental Railways.

[17]Mott, Between the Ocean and the Lakes, p. 44.

[18]Morton, Report. Kirkwood, Report, made similar arguments for his choice of broad gauge in Missouri.

[19]Economies of scale in production of rolling stock were exhausted at low enough levels that they did not affect choices among the more popular gauges. Indeed, variations among rolling stock designs were greater within the 4'8.5" gauge than between that gauge and any of the broader gauges, and stock for different gauges often varied only in the wheel trucks.

[20]In the other cases I have not been able to document the experience of the lines' chief engineers.

[21]Taylor and Neu, American Railroad Network, p. 35.

[22]Until at least 1871, Johnson (Transcontinental Railways) tried without success to establish the gauge as the standard in yet unbuilt western regions. Johnson is notable for the obstinacy with which he stuck with his preferred gauge, but his earlier career represents an exception to the rule that engineers stayed within a single engineering tradition. He assisted in building several 4'8.5"-gauge railways before proposing the 6'0" gauge.

[23]For extensive accounts, see Hilton, American Narrow-Gauge Railroads for the United States and Lavallee, Narrow-Gauge Railways for Canada.

[24]Mott, Between the Ocean and the Lakes, p. 45.

[25]Taylor and Neu, American Railroad Network, pp. 18-19.

[26]Ibid., p. 19.

[27]Seymour, Review, pp. 23-24. The choice was apparently influenced by the gauge's use on a short line in California. It is a splendid irony that the gauge of the secessionist southern states was thus favored.

[28]Taylor and Neu, American Railroad Network, pp. 55-56.

[29]Kirkwood, Report, also noted that eastern railways used several different gauges in any case.

[30]Reasons for this adoption are not well documented. Taylor and Neu, American Railroad Network, p. 44.

[31]Hilton, American Narrow-Gauge Railroads, p. 83.

[32]National Narrow-Gauge Railway Convention, Proceedings, p. 20.

[33]Hilton, American Narrow-Gauge Railroads, p. 91. Manufacturers' standardization of locomotive models, much more advanced than at the time of introduction of broader gauges, may also have helped to limit diversity among narrow gauges. Nevertheless, manufacturers could and did build to order, sometimes at a price premium.

[34]American Railroad Network, p. 44.

[35]From 1848 to 1852 a state law established 4'10" as the local standard gauge (ibid., p. 35). This law reflected the situation rather than created it, for railways that wished were able to adopt other gauges.

[36]Investments in the latter route by New York railways using both 4'8.5" and 6'0" led to adoption of 4'10" as a "neutral" gauge.

[37]Among these allies was E.F. Johnson's previously built Illinois and Wisconsin Railroad.

[38]Hilton, American Narrow-Gauge Railroads, p. 99.

[39]Hilton, American Narrow-Gauge Railroads, pp. 101-11.

[40]For some quotations see ibid., pp. 35, 65. Econometric tests would not shed much light on variations among gauges in cost and performance, both because data are sparse and noncomparable and due to systematic differences for the broad gauges in traffic and operating conditions, including transfer costs at breaks of gauge.

[41]Taylor and Neu (American Railroad Network) discuss cooperation extensively.

[42]Hilton, American Narrow-Gauge Railroads, pp. 37-8.

[43]Taylor and Neu, American Railroad Network, p. 80; Mobile and Ohio Railroad Co., Annual Report.

[44]Hilton, American Narrow-Gauge Railroads, p. 277.

[45]David, "Some New Standards," discusses gateway techniques and adapters in other contexts. Taylor and Neu, American Railroad Network, provide detail on their use on North American railroads.

[46]The final amount paid is unclear. Currie, Grand Trunk Railway, pp. 187-88.

[47]Lovett, Canada, p. 90.

[48]Pennsylvania Railroad Co., Annual Report.

[49]Taylor and Neu (American Railroad Network, pp. 81-82) note the wide use of 4’9” but have no account of the variant gauge’s origin in the resolution of a greater gauge difference.

[50]Farrell and Saloner, “Standardization.”

[51]Taylor and Neu, American Railroad Network, pp. 79-81.

[52]It is far from clear what effect the counterfactual alternative of government ownership would have had. Different U.S. states might well have adopted different gauges, as did the different Australian colonies, and might have found it more difficult to form interregional systems and arrange side payments to resolve diversity.

[53]Examples include Australia and Argentina, each with three different regional gauges; Spain and Portugal, with a different gauge than the rest of Europe; and countries of the former Russian and Soviet empires, with a different gauge than all other neighbors.

[54]An interesting task for future research would be to develop both upper- and lower-bound estimates of this inefficiency, although the well-known difficulties of counterfactual analysis make this an uncertain undertaking.

[55]David (“Clio”) does not emphasize the issue, but he does state that decentralization of agents may prevent efficiency-improving conversion of technique. Other economists cited by Liebowitz and Margolis (“Fable”) take market failure, however, as the main point of David’s paper. Liebowitz and Margolis (“Path Dependence”) argue that private profit opportunities will lead to conversion whenever it is socially efficient. I regard the present paper as a refutatation of their additional claim that the only sort of path dependence that challenges the “model of relentlessly rational behavior leading to efficient, and therefore predictable, outcomes” is the self-contradictory and thus non-existent sort in which an outcome is inefficient, remediable through profit-oriented behavior, but somehow not remedied.

[56]There are grounds to presume that market failure played a small role in the process, at least in the pace of conversion, but a full empirical test may require more data on costs and demand than are available, and such a test is in any case beyond the scope of this article. The private and social consequences of an individual railway line’s choice of gauge differed: a choice for a compatible gauge would benefit neighboring lines as well as itself. Presumably, then, railway lines took “too little” account of the effects of their choices on network integration. Side payments, appropriation, and cooperation internalized these externalities sufficiently to

produce socially optimal behavior for some railway lines, but not necessarily for all, and not necessarily as soon as optimal.

[57]"Uniformity in Rail-way Tracks," January 21, 1832, p. 51.

TABLE 1

INTRODUCTIONS OF GAUGES TO NEW REGIONS, 1830-1869[1]

Period[2]	Gauge	Region (number of separate locations, if more than one)
1830-1839	4’8.5”	Maryland, New York (4), Louisiana (3), Delaware, Pennsylvania (3), Virginia, Massachusetts, Alabama, Kentucky, Quebec, Maine, Ohio-Michigan, North Carolina, Indiana, Illinois
	4’10”	New Jersey, Ohio[3], Mississippi
	5’0”	South Carolina, Florida
	5’4”	Ohio
1840-1849	4’8.5”	Alabama (additional)
	5’6”	Maine-Quebec[3]
	6’0”	New York[3]
1850-1859	4’8.5”	Ohio[3,4] (additional), Texas
	5’0”	California
	5’6”	Ontario[3,4], Missouri, Texas[3,4], Louisiana[4]
	6’0”	Illinois[3], Ohio-Indiana-Illinois[3,4]
1860-1869	5’0”	Washington territory
	5’6”	Arkansas[4], Louisiana[4] (2 additional)

[1]Railway lines with no prior or simultaneously built neighbors of the same gauge. Unless otherwise specified, these lines had no neighbors of any gauge.

[2]Date of opening of first section of track.

[3]Railway line with prior neighbor(s) of different gauge(s).

[4]Railway line with apparent reason to anticipate a later common-gauge connection.

Sources: See text, note 5.