1   **How Can We Fix It If We Don't Know What's Broken?**
2   **Standardizing Processes and Tools to Evaluate Information Provided to Transit Riders**
3

4   **Authors**
5   Ruth Miller
6   Cal-ITP
7   ruth.miller@dot.ca.gov
8
9   Elizabeth Sall
10   UrbanLabs LLC
11   easall@gmail.com
12
13   Hunter Owens
14   Caltrans
15   hunter.owens@dot.ca.gov
16
17   Olivia Ramacier
18   Cal-ITP
19   Olivia.Ramacier@dot.ca.gov
20
21   Michael A. Chow
22   Jarvus Innovations
23   michael.c@jarv.us
24

25   **Word Count**
26   3853 words + (4 tables/figures x 250) = 4853 words
27

28   **Keywords**
29   Transit, transit data, GTFS, data, data quality, trip planning, journey planning, best practices,
30   validators, open data
31
32   **Submitted for TRB 2022**

1 **ABSTRACT**
2 128 words
3
4 Many transit riders rely upon third party applications to plan their transit trips, and these
5 applications in turn rely upon transit providers to publish high quality schedule information in
6 the General Transit Feed Specification (GTFS) format. GTFS data varies in terms of both
7 completeness and accuracy across transit providers. In order to support transit providers in
8 improving their data, Cal-ITP first needed to establish an assessment process. That innovative
9 process relies on five components: the GTFS reference specification, community-led GTFS Best
10 Practices, California's own minimum guidelines for GTFS quality, automated GTFS Quality
11 Reports, and a subjective GTFS Grading Scheme. Together, these components drive a transit
12 data assessment process that allows Cal-ITP to strategically allocate its resources to transit
13 providers and improve the GTFS data that reaches their riders.

1 **INTRODUCTION**
2 In under two decades, it has become commonplace for would-be transit riders to first look up
3 their journey on their smartphone. This shift in expectations means that standardized transit data
4 about available transit options is now an essential part of the transit journey-planning experience.
5 In most of the world, including North America, this transit data is represented in the General
6 Transit Feed Specification (GTFS) format (1). This data is only useful when it is correct and
7 provides a complete picture of what transit is available based on your individual travel needs.
8 Data which lacks key fields can render transit an unreliable option for swaths of people, and
9 incorrect data can damage a rider's transit experience more than no data at all (2).
10
11 The California Integrated Travel Program (Cal-ITP) is a new program of Caltrans' working in
12 cooperation with the Division of Rail and Mass Transit and others. Cal-ITP's vision is to
13 improve the transit customer experience from end-to-end through a set of targeted and strategic
14 actions, and improving GTFS data quality is one of its three initiatives. This paper documents the
15 key findings from Cal-ITP's research on how to measure transit data quality, and shares the
16 processes that Cal-ITP has designed and implemented through its Transit Data Quality Program
17 to regularly measure it consistently across the state to identify where California's 184 GTFS-
18 producing transit providers can improve in order to meet rider expectations and needs.
19
20 **BACKGROUND**
21 California has a large, disaggregated, and complex public transit ecosystem. There are over 300
22 public transit operators (depending on how one defines "public," "transit," and "operator")
23 including fixed-route and demand-responsive services, a dozen of different proprietary fare
24 payment systems, scores of mobile applications each for a limited service area, and various
25 regional entities responsible for setting policies related to transit. This disaggregation creates a
26 lack of standardization, creating friction and hampering the customer experience in unnecessary
27 and sometimes confusing ways.
28
29 The transit industry has designed its products and services to support the assumption that
30 customers stay within their transit provider's ecosystem and become relatively familiar with
31 it. However, commute patterns throughout California consistently traverse multiple transit
32 providers, necessitating a collective perspective of the entire customer experience. This
33 experience starts with discovery ("what options are available to me?"), and is most often
34 accomplished using a journey planning application on a smartphone or computer.
35
36 In most of the world, journey planning involves a transit provider publishing their planned
37 operations in the General Transit Feed Specification (GTFS) format, and this data then being
38 consumed by a third-party application (such as Google Maps, Transit App, etc.). These
39 applications allow users to optimize their journey, visualize the route, see estimated times of
40 departure, and generally understand what experience they will encounter. GTFS was originally
41 designed by Google Maps and Portland's TriMet in 2005 (3). Many extensions to the original
42 specification have been added over the years, including most notably GTFS Realtime in 2012.
43
44 Riders respond positively to the existence of trip planning data. Research shows that ridership
45 can increase two percent with the existence of real-time information (4, 5). While real-time data
46 is compelling for transit riders, it is unfortunately not widely available among California's transit

providers. As shown in Table 1, many California transit providers began 2020 either publishing
no standard digitized trip planning information.

Table 1 - Implementation of GTFS in California in early 2020

GTFS Availability	Number of Transit Providers	2019 Share of Ridership,California	2019 Share of Transit Revenue, California
No Published GTFS	89 (41.4%)	17.3%	17%
GTFS Schedule only	102 (47.4%)	45.1%	45%
GTFS Schedule and GTFS Realtime	24 (11.2%)	37.6%	38%

Source: Cal-ITP analysis done by Trillium Transit, February 25, 2020 using 2019 National
Transit Database (6)

Due to this modest publication of data in the GTFS Realtime standard, transit users are unable to
reliably access the real time transit information that they have come to expect and depend on
from other modes, such as driving and TNCs (7). The lack of information could decrease the
trust and satisfaction of current transit riders (8) and discourage new users from trying transit
(9).

METHODS
Cal-ITP held a market sounding event in fall 2019 with companies and organizations in the
payments and trip planning industry to identify barriers to seamless transit travel and identify
possible solutions. This event identified three initiatives.
- Ensure access to reliable and accurate real time transit information;
- Reduce friction in payments; and
- Create a statewide eligibility verification system.

This event was followed in early 2020 by thoroughly analyzing the feasibility of the potential
solutions and fully assessing their economic impact (10). Cal-ITP was able to confirm that all
three initiatives were feasible and would yield additional economic benefits under even
conservative to moderate assumptions of costs and ridership effects. The transit information
initiative moved forward with the decision to implement GTFS Realtime for all fixed route bus
and rail transportation services in the state, as well as other GTFS extensions as they are adopted.
Specifically, Cal-ITP proposed the following actions.
- Officiate GTFS as the statewide standard for transit data
- Expand GTFS to serve more use cases
- Develop a common GTFS infrastructure
- Initiate and support the creation of California Implementation Guidelines for GTFS
- Create and maintain a program to establish, incentivize and maintain compliance with the statewide standard including the development of California Implementation Guidelines
- Provide a way for local agencies to source GTFS implementation support

1   • Provide an ongoing way for local agencies to source equipment (such as Automated
2   Vehicle Location equipment) to be able to publish and communicate real-time
3   information and other aspects of mobility data standards as they are developed
4   • Provide a way for local agencies to source passenger-facing equipment (such as
5   information displays)
6
7 During spring and summer 2020, Cal-ITP held a series of interviews with a variety of
8 organizations to inform implementation of the above actions. These interviews included
9 conversations with:
10   • Third party trip planning applications
11   • Vendors of GTFS tools and support
12   • A variety of California transit agencies across urban and rural service areas through
13   recurring working groups
14   • Transit data managers in a variety of local and regional governments in California
15   • Other Departments of Transportation working with GTFS
16   • Non-profit advocates and academic researchers focused on GTFS
17
18 Informed by these discussions, Cal-ITP began creating its Transit Data Quality Program. This
19 program is intended to improve the trip planning experience for California's transit riders. The
20 first step to improving this data is understanding it, and the Transit Data Quality Program relies
21 upon five tools that are explained in the next section.
22
23 **DATA QUALITY TOOLS**
24 Cal-ITP uses the five tools below to assess the quality of GTFS data produced by California's
25 transit providers. Identifying opportunities for improvement is a critical first step for Cal-ITP's
26 Transit Data Quality Program.
27
28 *GTFS Reference Specification*
29 The format and structure of the files that comprise a GTFS dataset are defined in the reference
30 documentation at https://gtfs.mobilitydata.org. The documentation is hosted by Google and
31 managed as an open standard with community involvement facilitated by the non-profit
32 MobilityData organization. The specification can be changed with community support once at
33 least one transit data producer and one consumer are identified to implement the change.
34
35 Transit operations are complex, and GTFS serves the global transit ecosystem, so the
36 specification has grown and changed over time (including the addition of GTFS-Realtime in
37 2012). In Cal-ITP's research, it identified a readiness among both transit data producers and
38 consumers to support more fields contending with payments and accessibility concerns. Cal-ITP
39 leveraged California's market share to drive discussions with data consumers and committed to
40 producing data for a significant number of providers to drive adoption of several extensions.
41   • GTFS Fares version 2, including fare capping and accepted payment types
42   • GTFS Pathways
43   • GTFS Flex version 2
44   • GTFS Text-to-Speech
45   • wheelchair_boarding fields for both stops.txt and trips.txt
46

1   The adoption of these extensions ensures that enough information is available to help all transit
2   users confidently plan their journey.
3
4   ***GTFS Schedule Best Practices and GTFS Realtime Best Practices***
5   The GTFS Schedule spec reference documentation may be the legal code for GTFS Schedule,
6   but the GTFS Schedule Best Practices (11) are the case law that help data producers understand
7   how the specification should actually be applied. The GTFS Schedule Best Practices were first
8   published in 2017 and are managed today by MobilityData on behalf of the transit data
9   community following a community-based contribution and governance process.
10
11 To complement the GTFS Schedule Best Practices, MobilityData recently formally adopted the
12 corresponding GTFS Realtime Best Practices (12) developed by Sean Barbeau at the Center for
13 Urban Transportation Research. The addition of these Best Practices reduces ambiguity for
14 transit data producers and consumers, which will result in fewer dropped predictions and greater
15 estimation accuracy.
16
17 ***California's Minimum GTFS Guidelines***
18 Based on its market soundings, interviews, and transit working groups, Cal-ITP developed a
19 draft California Minimum GTFS Guidelines in summer 2020. Circulating this draft, hosting
20 public information sessions, and coordinating followup interviews gave Cal-ITP a better
21 understanding of the tradeoffs and reasoning behind each party's desires. Balancing these needs,
22 Cal-ITP authored and published version 1 of the California Minimum GTFS Guidelines in
23 September 2020 on the Caltrans website (13).
24
25 The Guidelines are composed of overarching Principles and a Data Process Checklist that
26 articulates the technical requirements to support the Principles and comply with the Guidelines.
27 The Data Process Checklist references the GTFS Reference Specification and GTFS Best
28 Practices to avoid placing additional burden on transit providers. The Guidelines are updated
29 annually in late summer with a similar release candidate process. The ten-point Data Process
30 Checklist from version 1 of the Guidelines is displayed in Figure 1.

1 **Figure 1 – The Data Process Checklist from Version 1 of the California Minimum GTFS**
2 **Guidelines**

In order to adhere to the above principles, public transit providers shall follow processes that are consistent with the following prioritized checklist:

1. Publish current transit data (GTFS and GTFS Realtime 2.0 feeds) at static fetch URLs.
2. Publish static GTFS data feeds that include a high-fidelity shapes.txt, persistent agency, stop and route identifiers, and implement all recommendations made in the Best Practices for GTFS for transit services open to the general public for which the provider is responsible.
3. Describe planned service changes in the static GTFS feed when they are known in advance. Publish a static GTFS feed update at least two weeks ahead of service updates to provide time for engaging with feedback from developers and data feed consumers.
4. Publish GTFS Realtime 2.0 Vehicle Positions, Trip Updates, and Service Alert feeds for fixed routes in the provider's officially published static GTFS data feed. Real-time vehicle information should be updated every 20 seconds or faster, have a trip_id that matches the static GTFS feed, persistent unique vehicle ids, and per-vehicle timestamps.
5. Make their GTFS and GTFS Realtime data feeds available on the feed aggregators https://transit.land and openmobilitydata.org.
6. Publish static GTFS pathways (pathways.txt, levels.txt) for infrastructure operated by the provider with stairs, escalators, elevators, or other accessibility considerations.
7. Offer GTFS and GTFS Realtime data without an extensive legal agreement, preferably under an open data license.
8. GTFS Realtime APIs should have an uptime of greater than 99%.
9. Publish a single point of contact for GTFS data on the provider's website, and in the feed info component of GTFS, as well as offer a way for data consumers to register with the transit provider so that they can be easily reached.
10. Maintain a process for tracking and improving accuracy of technology systems and revisiting best practices and these Guidelines.

3 *Source: Caltrans Website, July 22, 2021 https://dot.ca.gov/cal-itp/california-minimum-general-*
4 *transit-feed-specification-gtfs-guidelines*
5
6 *Automated GTFS Quality Reports*
7 Next, Cal-ITP created tools to assess the quality of data against the Guidelines. Much of the
8 Guidelines can be assessed programmatically, such as the availability of certain files/fields such
9 as shapes.txt. MobilityData also maintains a canonical GTFS validator, which allows anyone to
10 replicate the quality assurance process at Google Maps and elsewhere.
11
12 Cal-ITP built a database to download and archive all known GTFS Schedule and GTFS Realtime
13 feeds in the State of California. They deployed a local version of the GTFS validator, running it
14 nightly and storing the results for manual query by Cal-ITP staff. Next, Cal-ITP designed a set of
15 stylized retrospective reports, and in June 2021 began generating one for each feed at the end of
16 each calendar month. These reports are available in a publicly-accessible archive (14) and
17 emailed directly to known contacts at each transit provider. Vendors are encouraged to subscribe
18 to updates for their clients, as well. An example of an automated report is shown in Figure 2.

1   **Figure 2 – Example Monthly GTFS Quality Report**

2
3   *Source: https://reports.calitp.org, July 22, 2021*
4
5   ***GTFS Grading Scheme***
6   Other components of the Guidelines, and quality more generally, are too subjective to
7   programmatically assess and require human judgement. MobilityData also developed and
8   maintains a methodology called the GTFS Grading Scheme (15), which directs a reviewer to
9   sample various fields and compare their contents to other sources published by the transit
10   provider. For example, a route may be called "Truckee/Northstar Evening" on printed maps but
11   "NSPM" in the GTFS feed, and that inconsistency could confuse a rider.

1   The Grading Scheme exists as a document describing the methodology, but earlier Caltrans users
2   found it too confusing to implement. With the help of Ian Wesley at the Washington State
3   Department of Transportation, Cal-ITP was able to develop a friendly interface that guides users
4   through the grading process. The tool randomly selects sample data, displays the sample data
5   alongside the grading criteria, and even generates hexadecimal colors (route colors) and aerial
6   maps (stop location) where helpful. The user's grades are stored in a report that can be easily
7   referenced, exported, and shared. A screenshot of the stop location question is shown in Figure 3.
8
9   **Figure 3 – Screenshot of Grading Scheme User Interface**
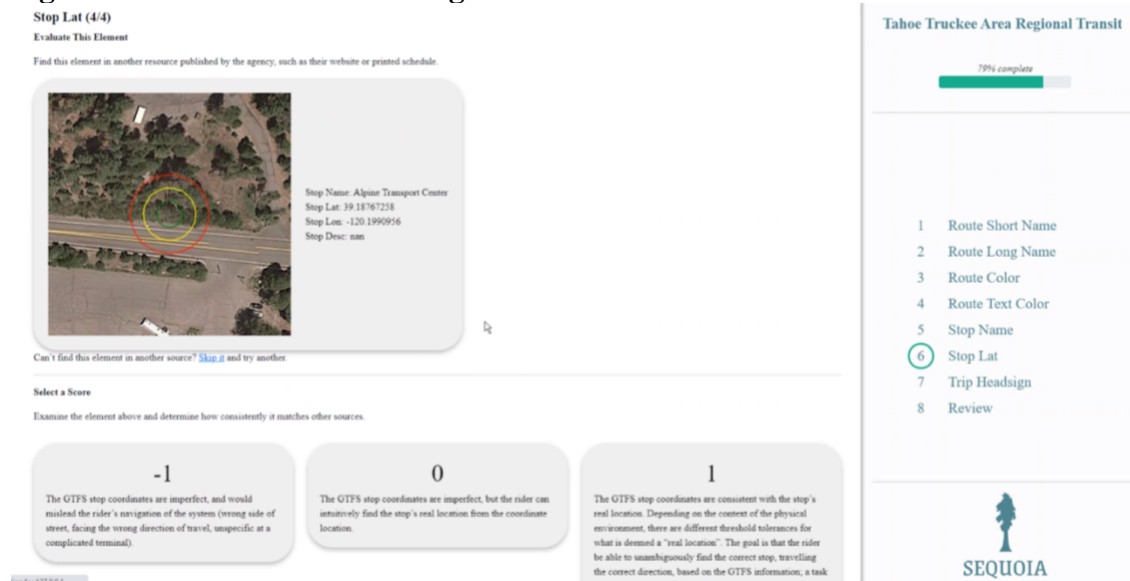
10
11   *Source: Washington State Department of Transportation, February 24, 2021*
12
13   **PROCESS**
14   With the above Data Quality Tools in place, Cal-ITP embarked on a process to apply the tools.
15   The first Transit Data Assessments were created manually in October 2020, before either the
16   queryable Validation results or the Grading Scheme interface were in place. Cal-ITP created a
17   template Transit Data Assessment, that allowed a reviewer to comment on how well a transit
18   provider met all ten of the items in the Guidelines' Data Process Checklist. During this early
19   process, Cal-ITP staff wrote thoughtful, individually crafted explanations specific to each transit
20   provider. Transit providers were reviewed based on their application status for various sources of
21   state funding, and the reports were delivered to transit providers by Cal-ITP staff.
22
23   As the assessments continued, the process was streamlined. The questions in the report were
24   modified into simpler "yes or no" formats. Since the Validator was deployed, the results are
25   pasted directly into the Transit Data Assessment report. Caltrans staff, who have strong existing
26   relationships with each transit provider, deliver to the transit provider along with a Doodle poll to
27   schedule a hand off presentation and discussion. Once the more simplified process was deemed
28   stable, additional staff from the Caltrans were trained to create and present the reports,
29   amplifying the effort. The assessment schedule was also made more predictable for agencies
30   based on the 12 Caltrans districts (District 1 reviewed in January, District 2 in February, etc.).

1   During the presentations and discussions with Caltrans, transit providers are able to understand
2   their GTFS data quality at a deeper level, often for the first time. Caltrans is able to offer a
3   variety of interventions to resolve these quality issues, including free GTFS data creation, low-
4   cost Computer-Aided Dispatch / Automatic Vehicle Location (CAD/AVL) hardware, Cal-ITP's
5   free GTFS helpdesk, and leveraged statewide procurements (16). Once transit providers and
6   Caltrans agree on next steps, those are written into a multi-year Transit Data Improvement
7   Strategy, which Caltrans keeps on file to ensure both parties are meeting their obligations to
8   transit providers and the data-consuming public.
9
10   **RESULTS**
11   Between the midpoints of 2020 and 2021, Cal-ITP's Transit Data Quality Program has assessed
12   194 California transit providers. This process identified the following gaps. Of all California
13   transit providers that publish GTFS data, at the start of 2021:
14   • 100% don't publish pathways data
15   • 81.4% do not specify terms of use for their GTFS data
16   • 42.9% don't mention their GTFS data on their website
17   • 14.9% don't have a feed_info.txt file
18   • 9.5% have feed_start_date values more than a year in the past, and 5.6% have
19   feed_end_date values more than a year into the future, indicating that minor schedule
20   changes, such as cancelled trips or closed stops, are not being represented.
21
22   Cal-ITP found the following specific validation errors in GTFS Schedule feeds.
23   • 31 agencies publish distances from origin in their shapes.txt file that are the same or less
24   than their preceding stop.
25   • 15 agencies publish trailing or leading whitespaces before stop or route names.
26   • Six agencies publish stop times that do not increase across the stop sequence.
27   • Four agencies have duplicate routes with the same name.
28
29   Beginning in June 2021, Caltrans staff began having conversations with transit providers about
30   their assessments. Each of these discussions culminates in the creation of a Transit Data
31   Improvement Strategy. Each strategy outlines the gaps between the provider's current GTFS data
32   and the steps the provider and Caltrans will take to close those gaps. Seven strategies have been
33   developed by the end of July. Of these:
34   • 100% will add language to their website indicating open usage terms for their GTFS data
35   • 85.7% will add direct links from their website to their GTFS data
36   • 42.5% will work with Caltrans to create and publish GTFS Fare data.
37
38   In parallel, Cal-ITP's helpdesk was able to help transit providers implement the following
39   improvements.
40   • Transit providers without GTFS Schedule data decreased from 89 to 17.
41   • Transit providers with GTFS Fare version 2 data increased from 0 to 17.
42   • Seven transit providers are working with Cal-ITP to deploy GTFS-Realtime for the first
43   time.

CONCLUSION

Cal-ITP and its partners have developed a suite of tools and a process that support the assessment of GTFS data quality. California, and others, have an opportunity to improve the completeness and accuracy of trip planning information, which would increase trust in the transit riding public, improve ridership, and support essential climate and equity goals.

Data quality, and fundamentally rider facing information, is not a one and done, lift and shift-style opera; it's an operational, ongoing need that requires commitment from agencies, vendors, and regulators. Steps must be taken to quantify and understand each step in the scheduling and data production workflows.

Without tools such as these, it is impossible to quantify the extent of bad data and debug the process that creates inaccurate rider facing information. Focusing on GTFS data as a first priority for accuracy and completeness makes sense in a mobile-first world, where a majority of riders are using cell phones to plan and execute trips.

Finally, enhanced GTFS data quality significantly opens up other opportunities for improving operational performance, reducing reporting burden on agencies, and introducing contactless fare payment systems, easier-to-use scheduling software, and interoperability with other internal systems.

FURTHER RESEARCH

Cal-ITP's research thus far has focused on GTFS Schedule data, so it follows that future research should focus on extensions, namely Realtime, Fares, Pathways, and Flex.

In fall 2021, Cal-ITP will add results from the CUTR GTFS-Realtime validator to its monthly reports. As the conversation about realtime data quality intensifies, the transit data community would be well served to have reached consensus on the following points.

- What constitutes "availability" of real time data? What is an appropriate denominator: vehicles scheduled, vehicles in service, or vehicles emitting real time locations?
- Is it reasonable, with existing technology, to expect 100% of active trips to appear in a real time feed? What would be required to justify this base expectation?

Based on preliminary discussions about realtime data quality, the need for solutions to the following questions is evident.

- How can the transit industry as a whole achieve matched Trip IDs? Trip IDs should match between schedule and realtime data, as defined in the best practices for both standards, but this is often not the case due to technical or process limitations.
- How should minor planned service changes be represented in data? For example, a stop closing for a week due to construction. Should this be published in the GTFS Schedule data (if so, how far in advance), the GTFS-Alerts, or GTFS-ServiceChanges? What method(s) are most likely to reach the rider through trip planning applications?

As new extensions are adopted and implemented (17), they will benefit from their own best practices and validation tools. This applies to:

- GTFS-Fares v2
- GTFS-Pathways v1
- GTFS-Flex v2.1

WORKS CITED

1. MobilityData. 2020. "GTFS Background." GTFS Resource Center. https://gtfs.mobilitydata.org/about.
2. Carrel, Andre, Anne Halvorsen, and Joan L. Walker. 2013. "Passengers' Perception of and Behavioral Adaptation to Unreliability in Public Transportation." *Transportation Research Record* 2351, no. 1 (January): 153-162. 10.3141/2351-17.
3. McHugh, Bibiana. 2013. "Pioneering Open Data Standards: The GTFS Story." Beyond Transparency. https://beyondtransparency.org/chapters/part-2/pioneering-open-data-standards-the-gtfs-story/.
4. Tang, Lei, and Piyushimita (Vonu) Thakuriah. 2012. "Ridership effects of real-time bus information system: A case study in the City of Chicago." *Transportation Research Part C: Emerging Technologies* 22, no. June 2012 (June): 146-161. 10.1016/j.trc.2012.01.001.
5. Brakewood, Candace, Kari Watkins, and Greg MacFarlane. 2015. "The impact of real-time information on bus ridership in New York City." *Transportation Research Part C: Emerging Technologies* 53, no. April 2015 (April): 59-75. 10.1016/j.trc.2015.01.021.
6. Federal Transit Administration. n.d. "National Transit Database." https://www.transit.dot.gov/ntd/ntd-data.
7. Crane, Jackson and Greg Rucks. 2017. "A consortium approach to transit data interoperability." *Rocky Mountain Institute*. https://rmi.org/wp-content/uploads/2017/03/consortium_approach_to_ITD_report2016.pdf
8. Zhang, Feng, Qing Shen, and Kelly J. Clifton. 2008. "Examination of Traveler Responses to Real-Time Information about Bus Arrivals using Panel Data." *Transportation Research Record* 2082, no. 1 (January): 107-115. 10.3141/2082-13.
9. Brakewood, Candace, and Kari Watkins. 2018. "A literature review of the passenger benefits of real-time transit information." *Transport Review* 39, no. 3 (May): 327-356. 0.1080/01441647.2018.1472147.
10. California Integrated Travel Project. 2020. "Cal-ITP Feasibility Study." Caltrans Website. https://dot.ca.gov/-/media/dot-media/cal-itp/documents/calitp-feasibility-study-042420-a11y.pdf.
11. MobilityData. 2021. "GTFS Best Practices." GTFS Resource Center. https://gtfs.mobilitydata.org/best-practices/.
12. MobilityData. 2021. "GTFS Realtime Best Practice." GTFS Resource Center. https://gtfs.mobilitydata.org/best-practices/gtfs-realtime

13. Caltrans. 2021. "California Minimum GTFS Guidelines." https://dot.ca.gov/cal-itp/california-minimum-general-transit-feed-specification-gtfs-guidelines.

14. California Integrated Travel Project. 2021. California GTFS Quality Dashboard. https://reports.calitp.org/.

15. MobilityData. 2020. "GTFS Grading Scheme." https://github.com/MobilityData/gtfs-grading-scheme.

16. California Integrated Travel Project. 2020. California Mobility Marketplace. https://www.camobilitymarketplace.org.

17. MobilityData. 2021. GTFS Extensions Roadmap. http://bit.ly/gtfs-extensions.