Discussion-piece #16
Calibration and Validation of Travel Models for New Starts Forecasting
Federal Transit Administration
June 6, 2006

1. Motivations. The calibration and validation of travel models provide the best evidence that the models grasp the transit supply characteristics and traveler behaviors that are crucial to subsequent forecasts for New Starts projects. Consequently, both project sponsors and FTA have a substantial interest in appropriate efforts to calibrate and validate models that provide a solid basis for predictions of the ridership and mobility benefits of major transit projects. FTA guidance on calibration and validation will outline FTA's priorities in the review of the adequacy of local travel models to support New Starts forecasts.

2. Traditional model development. While there is no universal approach to model development, the process is often described as comprising three phases:

   - Estimation – where individual records from survey data are used with specialized software to estimate the values of model parameters and draw conclusions on the appropriate variables and structure for component models;

   - Calibration – where each component model emerging from estimation is implemented in its application software and adjusted to reproduce current travel behaviors; and

   - Validation – where the entire model set used to make a "forecast" of current travel patterns and demonstrate sufficient ability to reproduce highway counts and transit line volumes.

   Several years of observations on the performance of models in New Starts forecasting suggests that this traditional approach – at least as it is commonly employed – has some shortcomings as preparation for transit forecasting.

   a) Misallocation of effort. Many model-development efforts reserve relatively few resources for calibration and validation – and those resources are often eroded by higher-than-anticipated efforts in data assembly and model estimation. Model estimation is an important step: it provides the basic parameters for all components of the model set, it has provided all of the insights into travel behavior available to the practice of travel forecasting, and it is the avenue to all future improvements in travel models. However, model estimation can easily consume all resources available for model development. The pursuit of clean estimation results (with all parameters estimates statistically significant and with the correct sign, for example) can extend the effort beyond allocated resources. The substantial penalty for that outcome is the reduction – perhaps elimination – of resources for calibration and validation. These latter phases of model development focus directly on the model set itself and its ability to produce useful forecasts – the crucial property for a model set that is about to be used to make forecasts that, for New Starts projects, may cost hundreds of millions of dollars or more. The failure to reserve sufficient resources for calibration and validation is a potentially severe strategic error in the preparation of reliable travel forecasting procedures.

b) Insufficient data. Useful forecasts rely on the ability of travel models to make reasonable estimates of a range of characteristics of the transportation system (travel speeds, transit accessibility, and others) – and the travelers who use the system (choices of destination, travel mode, transit-access mode, transit path, and many others). Remarkably, many model-development efforts proceed with only limited data to verify the estimates produced by the models. In particular, the absence of insights on transit ridership patterns – that are available only from carefully designed surveys of current transit riders – is often an alarming gap in the foundation of models prepared to make ridership forecasts for major transit investments.

c) Superficial calibration. Perhaps the largest problem for transit forecasting that occurs in traditional model development is a transit calibration effort that results in adjustments necessary to match current data that are no more than correction factors for errors made elsewhere in the model set. The "calibration" of alternative-specific constants is meaningful only when the person-trip tables, highway and transit networks, and observed transit ridership patterns are sufficiently accurate. Errors in person-trip tables, in particular, have frequently led to grossly distorted calibration constants that have nothing to do with travel behavior and that lead to useless transit forecasts.

d) Inattention to forecasting properties. The traditional focus on matching current conditions in testing a new model set is necessary: models that cannot describe current conditions offer little promise for predicting the future. However, travel models are ultimately used to make predictions of changes – between today and the future and (particularly in New Starts forecasting) between alternatives in the future. Inattention to the forecasting properties of models can permit the introduction of model properties that produce inexplicable forecasts. Bizarre alternative-specific constants for transit line-haul modes are a common outcome of the exclusive focus on matching current conditions without regard to their implications in forecasting. Arbitrary decision rules have the same perverse effect. For example, a common "remedy" in model calibration for the overestimation of short transit trips has been to reclassify any transit trip of less than N minutes in-vehicle time as an auto trip. With the correct value of N, that rule can produce a better match against current transit trip lengths. In forecasting, however, it can produce havoc as transit guideways shorten the travel time for some trips sufficiently to trigger the rule – and thereby predict a loss of transit riders for a faster guideway alternative compared to a slower local-bus baseline conditions.

e) Too-late initial forecasts. Forecasts with new models often tell as much about the models as the transportation alternatives to which they are applied. The differences between the forecasts – today versus the future and across changes in the transportation system – illuminate the forecasting properties of the models better than any cross-sectional testing done during "base-year" calibration and validation. While helpful, the late discovery of previously unknown model "properties" is very awkward when the properties are bizarre, the forecasts are inexplicable, and various studies face milestones that depend on those forecasts.

3. Some improvements on current practice. The remedy for these problems lies in sufficient data collection, meaningful calibration, useful validation, and informative documentation.

   a) Sufficient data collection. Models intended to support decisions on major transit investments should be based on data that are sufficient to support the reliability of the models. Four kinds of data are essential to calibration of the transit-forecasting capabilities of a model set, and a fifth is highly desirable:

      - Highway speeds – because of their roles in (1) predicting transit speeds, (2) representing impedances in trip distribution, and (3) representing the principal alternative to transit in mode choice;
      - CTPP2000 – the Census Transportation Planning Package – because large errors in work travel-patterns can substantially distort mode-choice models for work trips;
      - Bus travel times – because the principal way that transit guideways improve transit service is to increase transit speeds, so comparisons between guideway travel times and bus travel times in mixed traffic simply must describe meaningful differences;
      - Transit rider survey – because transit line counts and on-off counts, by themselves, cannot inform travel models sufficiently to provide a solid basis for development of the transit components of a model set; and
      - (Ideally) a household diary survey – because non-work travel patterns are essential to the development of useful non-work mode choice models.

   b) Meaningful calibration. In contrast to the often superficial traditional calibration (particularly of mode-choice models), a meaningful calibration effort ought to exhaust the available data. Virtually any mode choice model can be made to match total transit ridership – even subtotals by access mode and socio-economic class – with sufficient hammering on its alternative-specific constants. Rigorous calibration would include:

      - Careful comparison of point-to-point travel times via the highway and transit networks;
      - Detailed inspection of person-trip tables – ***not*** simply trip-length frequency distributions – in terms of district-to-district interchanges; this inspection is particularly important for significant transit markets by demonstrating, for example, that the distribution model predicts reasonably well the production locations of person-trip attracted to the central business district and other transit-competitive attraction areas;
      - The assignment of "observed" transit trip tables (derived from a transit rider survey) to the coded transit networks to isolate the network coding and transit pathbuilding components of a model set; this test permits comparisons of individual path characteristics, transfer frequencies, boarding/alighting volumes at

stations, trip lengths, and the geographic locations of trip productions and attractions by access mode and socio-economic class; and it supports adjustments to network coding rules and pathbuilding parameters, and often highlights problems in person-trip tables;

- The comparison of transit trip tables from the mode choice models against the "observed" patterns derived from a transit rider survey; this test examines the extent to which the model set grasps the nature and magnitude of the principal ridership markets – by trip purpose, socio-economic class, transit access mode, and perhaps other characteristics; and
- The comparison of transit volumes on individual transit lines (or groups of lines in the case of local buses), guideway facilities, stations, and park/ride lots – and between station pairs if the data are available; this test demonstrates the ability of the model set to replicate loadings on the principal components of the transit system.

c) Useful validation. Because rigorous calibration exhausts the available data, no data comparisons remain for validation. Two important tests remain, however.

- Interpretation of the story told by the models themselves – particularly transit network coding, transit pathbuilding, and the mode choice models – about the behavior of travelers; this test helps to ensure that the various parameters, constants, network coding conventions, and other decision rules in the models tell a coherent story about travel behavior; it prevents (by highlighting the need for correction) implausible relationships (values of time, ratios between weights on impedance variables, internal inconsistencies) and explains the properties of the models to non-travel-forecasters; and
- Demonstration of reasonable predictions of change between today and the future and in response to changes in the transportation (particularly transit) system; this last set of tests adds a major new dimension to the understanding of the properties of a new model set: the ability to respond reasonably to demographic growth and consequent changes in congestion and parking costs, and to produce coherent responses to major changes in the transit system. (To be useful, tests of reactions to change must be done through applications of the model in full production mode. Consequently, simple elasticity tests – done by rescaling impedance matrices and then rerunning the mode choice model – are insufficient because they do not exercise the full range of model components, particularly network coding conventions and transit pathbuilding parameters that are central to the transit-related properties of a model set.

d) Informative documentation. Rigorous calibration and validation produces a rich set of quantitative results. Beyond presenting those results, the most compelling documentation of the readiness of a model set for transit forecasting lies in the presentation of:

- The significant travel markets that exist in the current transit system;
- The ability of the model set to describe the nature and magnitude of those markets;

- The reasonableness of predicted changes in those markets in response to changes in model inputs; and
- The limitations imposed on the model set by the current markets (choice riders), behaviors (significant park/ride usage), and/or transit modes (fixed guideways) that will require special attention and introduce relatively greater risks in forecasts that rely on those currently missing conditions.

4. <u>Implementation.</u> The purpose of FTA guidance on travel models is to describe the way that FTA reviews models in response to the requirement that the agency consider the reliability of forecasts in rating proposed New Starts projects. Because guidance is not regulation, it conveys no mandates to project sponsors. Consequently, project sponsors are not required to respond specifically to any of the principles outlined here and forthcoming in formal guidance. The single exception may be a requirement for recent and useful on-board survey data, depending on the New Starts provisions in the final regulation implementing SAFETEA-LU. That proposed requirement reflects the singular importance of data on current transit ridership as a basis for both informing travel models and underpinning the case for a major transit investment.

   Good practice in model development has positive consequences in New Starts project development, however. Forecasts from models that have been rigorously calibrated and validated are likely to generate fewer problems with acceptance by FTA for project evaluation and will contribute to more favorable characterizations of their inherent risks. Where models have been prepared through less rigorous efforts, the absence of fewer compelling arguments may lead to follow-up testing of model properties – particularly when the New Starts forecasts include unexpected characteristics – and the possible need for corrections to the models. Ultimately, less favorable characterizations of risk are likely outcomes of less-well-supported models and forecasts.