KENNETH A. SMALL
University of California, Irvine

CLIFFORD WINSTON
Brookings Institution

JIA YAN
Hong Kong Polytechnic University

Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design

THE U.S. HIGHWAY SYSTEM, largely constructed with public funds from the federal road user tax, could be characterized as a public good if it were rarely congested. But like many public goods that are available at little or no charge, its quality has deteriorated with the intensity of use. Today, the nation's road system has turned into a "tragedy of the commons" as road users experience nearly 4 billion hours of annual delay.[1] Of course, even an efficient road system would force motorists to incur some delays, but the current level is regarded by most observers as excessive.

Historically, the public has had a *status quo* bias against economists' recommendations to use the price mechanism to reduce congestion.[2] Policymakers therefore have pursued other approaches, such as allocating reserved lanes to vehicles carrying two or more people. But recent evidence indicates that these *high-occupancy vehicle* (HOV) lanes sometimes carry fewer people than general-purpose lanes, attract many family members who would ride together anyhow, and shift some travelers from vanpools or buses to low-occupancy carpools.[3] As a result, HOV lanes are losing favor among state transportation departments.

1. Schrank and Lomax (2005).
2. Small, Winston, and Evans (1989); Mohring (1999). The papers in Santos (2004) provide recent discussions of road pricing.
3. Orski (2001); Poole and Balaker (2005).

A recent innovation is to fill the reserve capacity not used by HOVs with solo drivers willing to pay a toll. These so-called *high-occupancy toll* (HOT) lanes can be found in the Los Angeles, San Diego, Houston, and Minneapolis–St. Paul metropolitan areas, and they are currently under consideration in other cities including Denver, Seattle, San Francisco, and Washington.

HOT lanes appeal to a broad set of motorists who are sufficiently inconvenienced by congestion to pay a sizable toll to travel on less-congested lanes, either daily or as dictated by their schedules. Although the adoption of HOT lanes in some urban areas indicates that the public is no longer opposed to all forms of congestion pricing, HOT lanes are questionable on welfare grounds for two reasons. First, motorists continue to impose high congestion costs on each other because most of the highway is unpriced. Second, the express lanes are still underused because a big price differential exists between the two roadways.[4] Indeed, simulations show that HOT lanes sometimes lower welfare compared with keeping all lanes in general use, particularly if they are priced high enough to allow motorists to travel at approximately free-flow speeds—a condition that is achieved to promote the service advantages of the lanes among the public.

In short, HOV and HOT policies do not appear to have answered the long-standing call for efficient yet politically viable road pricing policies. In this paper we seek to identify such policies by analyzing the behavior of motorists traveling on California State Route 91 (SR91) in Orange County. These travelers have the option of traveling solo on the general lanes, paying a toll to use the HOT express lanes, or forming a carpool to use the express lanes at a discount. Because travelers are likely to vary in their preferences for speedy and reliable travel, we model the situation accounting for both observed and unobserved preference heterogeneity.[5] We find that users of SR91 have high average values of travel time and travel-time reliability, and that the distributions of these values exhibit considerable dispersion.

We show that by designing differentiated pricing schemes for general and express lanes that cater to such varying preferences, it is possible to capture some of the efficiency that HOV and HOT policies sacrifice while generating welfare disparities among road users that are not only smaller than more-

4. Small and Yan (2001).

5. Previous empirical studies that allow for heterogeneous preferences among motorists include Calfee, Winston, and Stempski (2001); Hensher (2001); Jiang and Morikawa (2004); Steimetz and Brownstone (2005); Hess, Bierlaire, and Polak (2005); Small, Winston, and Yan (2005a). Simulation studies incorporating heterogeneity to analyze pricing scenarios include Small and Yan (2001); Verhoef and Small (2004); De Palma and Lindsey (2004).

efficient pricing policies generate but small enough to be comparable to policies that have actually passed the test of political acceptability in a few urban areas.

Empirical Model of Travel Choices

California State Route 91 is a major limited-access expressway used heavily by long-distance commuters. A ten-mile stretch in Orange County includes four free lanes and two express lanes in each direction. Motorists who wish to use the express lanes must set up a financial account and carry an electronic transponder to pay a toll, which varies hourly according to a preset schedule. Carpools of three or more people could use the express lanes at a 50 percent discount at the time our surveys were conducted.[6] Unlike the regular lanes, the express lanes have no entrances or exits between their end points.

We assembled a data set from surveys of travelers on the corridor, describing the lane choices they actually make and choices they hypothetically would make in different circumstances. In an earlier paper using the same data, we modeled motorists' lane choice only.[7] Here we model three simultaneous decisions by motorists: 1) whether to acquire a transponder, which gives them the flexibility to use the express lanes whenever they desire; 2) whether to travel on the express or free lanes for the trip in question; and 3) how many people to travel with in their vehicle. We include transponder acquisition as a separate choice because it captures a legal requirement to use the express lanes. Whether a motorist decides to meet the requirement may depend on other characteristics than those explaining day-to-day travel choices, and may cause more persistence in travel behavior than would otherwise be the case. We include vehicle occupancy so that we can explore the effects of various policies that depend on it.

The three choices are assumed conditional on related choices including travel mode (car versus public transport), residential location, and time of day of travel. In our context, mode choice is unimportant because public transportation has a very small share of travelers on the corridor served by SR91. Residential location indeed may be important, but it is a longer-run response that introduces complexities that are difficult to capture in a tractable empirical model. We would like to model the choice of what time of day people travel

6. Starting in late May 2003, these carpools could travel for free except when traveling outbound on weekday afternoons (from 4:00 p.m. to 6:00 p.m.), at which time there is a 50 percent discount. Current information is available on the Orange County Transportation Authority's *91 Express Lanes* website (www.91expresslanes.com).

7. Small, Winston, and Yan (2005a).

but are unable to do so because we lack information on how congestion varies on roads that people use besides the SR91 study corridor. Later in this paper we describe some statistical tests that indicate that our results are not particularly sensitive to our assumption that travel occurs at a given time of day.

In the empirical analysis that follows, we combine data that describe motorists' actual decisions for their morning commute with data indicating hypothetical choices between the express and free lanes under varying travel conditions. This strategy permits us to extract information about the distribution of preferences that would otherwise be impossible to extricate from other random influences on behavior. The mechanism at work is that an individual who answered the hypothetical questions provides responses for up to eight different scenarios. Hence we can effectively infer that individual's preferences through his or her unique pattern of responses to trade-offs among cost, time, and reliability. We estimate common coefficients of the trade-offs that are shared among individuals and coefficients of the trade-offs that vary among individuals, enabling us to measure the key distributions of the value of time and reliability in our sample.

Formally, traveler n faces a choice whether to have a transponder (T) or not (N); whether to travel on a general (free) lane (G) or express lane (X); and whether to travel with one, two, or three people in the vehicle (where three means three or more). The three choice dimensions define $2 \times 2 \times 3 = 12$ alternatives, but only nine of them are available because a highway traveler must have a transponder to use the express lane, thereby eliminating combinations $NX1$, $NX2$, and $NX3$.

Following standard discrete-choice modeling, we specify the indirect utility of traveler n choosing an alternative j to be random:

$$\mathbf{U}_{jn} = \mathbf{X}_{jn}\boldsymbol{\beta}_n + \varepsilon_{jn}. \tag{1}$$

In equation (1), $\mathbf{X}_{jn}$ is a vector of attributes associated with alternative j including the toll, travel time, and reliability that apply to the traveler's trip; $\boldsymbol{\beta}_n$ is a vector of parameters that captures the traveler's preferences for those attributes; and ε_{jn} is an error term capturing unobserved influences. We measure preference heterogeneity by allowing parameter vector $\boldsymbol{\beta}_n$ to vary across individuals according to both observed characteristics and random (that is, unobserved) influences:

$$\boldsymbol{\beta}_n = \mathbf{W}_n\boldsymbol{\gamma} + \boldsymbol{\mu}_n. \tag{2}$$

In equation (2), $\mathbf{W}_n$ is a vector of explanatory variables relating to traveler n, while $\boldsymbol{\mu}_n$ is a vector of random variables; $\boldsymbol{\gamma}$ is a vector of parameters, to be estimated statistically, describing how preferences depend on observed characteristics. The random terms $\boldsymbol{\mu}_n$ are assumed to be independent normal random variables, with variances to be estimated. Thus the term $\mathbf{W}_n\boldsymbol{\gamma}$ describes observed heterogeneity and $\boldsymbol{\mu}_n$ describes unobserved heterogeneity in preferences toward travel characteristics.

If ε_{jn} in equation (1) were independently distributed according to identical extreme-value distributions, then equations (1) and (2) would constitute a conventional mixed-logit model where each choice probability can be expressed as a standard multinomial logit choice probability (conditional on $\boldsymbol{\beta}_n$), integrated over the distribution of $\boldsymbol{\mu}_n$ (which determines $\boldsymbol{\beta}_n$).[8] Our model is more complicated because we specify the structure of ε_{jn} to account for certain special features of the data. One is the decision structure inherent in our choice alternatives. Another, which we describe later, is that we merge our data from several sources.

As noted, the decision structure involves three choice dimensions. Thus it is unlikely that the alternative-specific preferences ε_{jn} for the nine permitted alternatives are independent of each other. Rather, a natural approach is to specify random preferences for groups of alternatives.[9] We let ε_{jn} include four distinct preferences: for a transponder (*T*), for the express lane (*X*), for a two-person carpool (*H2*), and for a three-person carpool (*H3*). Thus

$$\text{(3)} \qquad \varepsilon_{jn} = \Delta_j^T \nu_n^T + \Delta_j^X \nu_n^X + \Delta_j^{H2} \nu_n^{H2} + \Delta_j^{H3} \nu_n^{H3} + \eta_{jn},$$

where Δ_j^T denotes a dummy variable equal to one if alternative j is one of those characterized by a transponder, and so forth. The four variables ν_n^k are independent normal random variables, each with mean zero and a standard deviation σ^k to be estimated. (For parsimony, we impose $\sigma^{H2} = \sigma^{H3} \equiv \sigma^{HOV}$.) The remaining random terms, η_{jn} in equation (3), are assumed to be independent random variables (one for each alternative) with identical extreme-value distributions, just like in a logit model.

Naturally, one can expect to estimate the distribution of only a few of the many behavioral parameters contained in a model like this. We choose two key parameters, which means there are two components of $\boldsymbol{\mu}_n$ in equation (2) with nonzero variances. One (μ_n^{Time}) is part of the coefficient of travel time, while

8. Small and Winston (1999) and Train (2003) contain expositions of the logit and mixed-logit models. The (normalized) extreme-value distribution for a random variable ε is defined by the probability $\text{Prob}[\varepsilon \leq x] = \exp(e^{-x})$.

9. For a description, see Brownstone and Train (1999). An alternative approach would be to use a nested-logit specification. The approach here is more flexible, typically better behaved numerically, and easier to implement given that we are using mixed logit.

the other (μ_n^{Rel}) is part of the coefficient of (un)reliability. Their standard deviations, to be estimated, are denoted σ^{Time} and σ^{Rel}.

To summarize the model's stochastic part, we specify six independent normal random terms (νs and μs) with five distinct unknown standard deviations (σs) to be estimated.

We define the value of travel time (VOT) and value of reliability (VOR) for individual *n* as the ratios of marginal utilities of travel time and reliability, respectively, to the marginal utility of money cost. That is,

$$VOT_n = \frac{\beta_n^{Time}}{\beta_n^{Cost}}, \tag{4}$$

$$VOR_n = \frac{\beta_n^{Rel}}{\beta_n^{Cost}}, \tag{5}$$

where in equation (1)

β_n^{Time}: the coefficient of travel time,
β_n^{Rel}: the coefficient of travel time reliability,
β_n^{Cost}: the coefficient of toll.

These values depend on observables $\mathbf{W}_n$ and random components $\boldsymbol{\mu}_n$ through equation (2).

Data Set and Econometric Issues

We combine survey data from three samples of people traveling between 4:00 a.m. and 10:00 a.m. on the California State Route 91 corridor westbound who have the option of using the express lanes. We collected data over a ten-month period in 1999 and 2000. The first survey was a telephone questionnaire generating 435 observations pertaining to actual travel on a particular day, conducted by researchers at California Polytechnic State University at San Luis Obispo (CalPoly) with our participation.[10] Thus it consists of revealed preference (RP) data.

The second and third samples are from a two-stage mail survey collected by us through the Brookings Institution. The initial stage collected RP data from seventy-nine respondents on actual trips taken during a week of travel, while a follow-up stage presented to each respondent eight stated preference

10. Sullivan and others (2000).

(SP) scenarios.[11] In each SP scenario, the respondent was asked to choose between two otherwise identical routes with specified hypothetical tolls, travel times, and probabilities of delay.[12] The SP sample contains seventy-eight respondents who generated 610 observations, and fifty-four of these people also answered the RP questions.[13]

By constructing a sample that contains both RP and SP observations, we can overcome the main drawbacks of each type of data. The use of RP data is often hindered by strong correlations among travel cost, time, and reliability, whereas SP data raise concerns about whether the behavior exhibited in hypothetical situations applies to actual choices. By specifying some parameters to be identical and others different in the utility functions generating RP and SP choices, we can improve the precision in estimating common parameters (due to low correlations designed into the SP questions) while allowing for expected behavioral differences in other parameters.

Table 1 presents some statistics on socioeconomic variables and trip distance. The Brookings RP sample appears to represent well the population characteristics of the SR91 catchment area, tracking census information for the two relevant counties except for household income—which, naturally, is higher for our respondents because most of them are commuters.[14] We estimate the average wage rate to be $23 an hour.[15] The CalPoly sample has higher household incomes and shorter trip distances than the Brookings samples, evidently being drawn from a smaller and more affluent geographical area. These sampling differences should not affect our parameter estimates because our model includes income and trip distance as explanatory variables.

11. The Brookings RP sample actually contains information for all commuting trips made within the survey week, which could be treated as separate observations. However, 87 percent of the respondents made the same choice every day and nearly all of the others varied on only one day. So we simplify, with little information loss, by creating a binary response variable equal to one if the respondent chose the express lanes for half or more of the days reported. We tried variants of this response variable with virtually no changes in results.

12. An illustrative scenario is presented in the paper's appendix.

13. Detailed descriptions of the samples are presented in Small, Winston, and Yan (2005b).

14. Our sample's median annual income is $46,250, whereas the average incomes in the two counties where our respondents lived were $36,189 and $39,729 in 1995, as estimated by the Population Research Unit of the California Department of Finance.

15. Data from the U.S. Bureau of Labor Statistics for 2000 record the mean hourly wage rate by occupation for residents of Riverside and San Bernardino Counties. We combine the bureau's occupational categories into six groups that match our survey question about occupation, and assign to each person in our sample the average Bureau of Labor Statistics wage rate for that person's occupational group. We then add 10 percent to reflect the higher wages likely to be attracting these people to jobs that are relatively far away.

Table 1. Socioeconomic Variables and Trip Distance

	Value or fraction of sample		
Variable	*CalPoly-RP*	*Brookings-RP*	*Brookings-SP*
Age (years):			
< 30	0.11	0.12	0.10
30–50	0.62	0.62	0.64
Household income ($000/year):			
< $60, 000	0.38	0.83	0.83
> $60,000	0.62	0.17	0.17
Female dummy	0.32	0.37	0.37
Mean actual trip distance (miles)	34.2	44.8	42.6
Number of respondents	435	79	78
Number of observations	435	369	610

Source: Based on data from the Brookings and CalPoly surveys and authors' calculations.
RP = Revealed preference data.
SP = Stated preference data.

Table 2 presents the choice shares of the nine alternatives associated with each RP sample. We observe a difference among carpooling propensities between the CalPoly and Brookings samples, with many fewer carpools in the latter. To better understand the difference, the CalPoly sample is disaggregated into four subsamples representing different ways of finding respondents.[16] The random subsample was obtained by telephone interviews drawn randomly from lists of telephone exchanges in the relevant area. The other three CalPoly subsamples came ultimately from license plates observed on the highway and are therefore choice-based (some were purposely carpool-enriched). Even in the CalPoly random subsample, however, the combined carpool shares are considerably higher (24 percent) than in the Brookings RP sample (6 percent), despite both being obtained from random telephone calls. The CalPoly random shares are much closer to the observed peak-period carpool shares on the SR91 roadway,[17] so we conclude that the Brookings sample undersampled people who carpool—possibly because the telephone screening questions to determine eligibility for the survey were originally designed to only find solo drivers and subsequently modified. Thus we use the CalPoly

16. Sullivan and others (2000).

17. Unfortunately, these are not measured at all precisely. Sullivan and others (2000) report the share of vehicles in a two-hour afternoon peak (eastbound direction) in 1999 to be 20.2 percent HOV2 and 3.7 percent HOV3. This would imply a share of commuters choosing to carpool of somewhat more than 24 percent, although how much more depends on how many of the passengers in carpools are also commuters. Also, because we model a four-hour morning peak period in the westbound direction, the comparison is not precise.

Table 2. Choice Shares of CalPoly and Brookings RP Samples
Percent

	CalPoly sample				
Alternative[a]	*Random*	*New plates*	*Repeat*	*UCI*	*Brookings RP sample*
*NG*1	41	28	17	11	51
*TG*1	16	26	33	39	23
*TX*1	19	15	16	22	20
*NG*2	7	9	3	0	0
*TG*2	3	8	16	6	2.5
*TX*2	8	5	7	22	1
*NG*3	3	3	0	0	0
*TG*3	1	3	3	0	0
*TX*3	2	3	5	0	2.5
All carpool	24	31	34	28	6
No. of observations	201	191	58	18	79

Source: Based on data from the Brookings and CalPoly surveys and authors' calculations.
RP = Revealed preference data.
UCI = University of California, Irvine.
a. Transponder acquisition: N = no, T = yes. Lane: G = general (free) lane, X = express lane.
Car occupancy: 1 = solo, 2 = HOV2, 3 = HOV3+.

random subsample as our measure of the population choice shares, and correct for choice-based sampling in our estimates by applying carpool-share weights to the other subsamples.[18]

We recognize that SR91 has a higher share of carpoolers than is the case for most other highways in the United States. Later in this paper we perform sensitivity analysis on the share of carpoolers to explore how our analysis applies to other U.S. metropolitan areas.

Specification and Estimation

We posit that motorists' joint choices are influenced by their socioeconomic characteristics and the characteristics of their journey, including the total trip distance and the toll, travel time, and (un)reliability of travel time on the portion of the journey where a lane choice exists.

The express lane toll for a given trip is the published toll for the time of day the motorist reported passing the sign that indicates the toll level. It is discounted by 50 percent if the trip was in a carpool of three or more. (We asked respondents, even in the SP survey, to indicate their vehicle occupancy for actual trips.)

18. Manski and Lerman (1977).

Our ability to measure individuals' preferences critically depends on capturing the different conditions they face when traveling at different times of day. Therefore we sought to measure those conditions carefully to construct variables for the RP portion of the analysis. We measured the reliability of service encountered by a traveler, as well as the travel time itself, by taking field measurements at many different times on eleven different days, corresponding approximately to the travel periods covered by our surveys. The field measurements consist of travel times clocked by students attending the University of California, Irvine, who drove the road repeatedly. Thus we were able to measure the median travel time observed across the eleven days, at any given time of day, as well as the entire distribution of travel times across those days, again as a function of the time of day. For our travel-time variable, whose coefficient is β^{Time} in equation (4), we use the median value. For our measure of unreliability of travel time, denoted *Rel* in equation (5), we use the difference between the 80th and 50th percentiles of the distribution of travel times across days. This measure captures the behavioral notion that people are more concerned with unexpected late arrivals than early arrivals. The measure also is less closely correlated with median travel time than a symmetric measure of dispersion such as the variance.[19]

The variables describing the individual include age, sex, household size, per capita income, total trip distance, and trip purpose (that is, a dummy for work trip). We explored other variables describing arrival-time flexibility, occupation, education, and workplace size, but found that they have little explanatory power and that omitting them did not materially influence the other parameter estimates.

The variables used in the SP analysis are defined, with one exception, identically to those in the RP data set, although the travel descriptors, of course, are generated differently, being specified in the survey questions instead of measured in the field. The one exception is reliability. We did not think we could explain percentiles of a probability distribution to survey respondents, so in the SP scenarios we described reliability as the frequency of being delayed ten minutes or more. We convert the responses into probabilities for purposes of analysis. The reliability variable therefore has different units and meaning in the RP and SP scenarios, so distinct RP and SP coefficients for it are estimated. However, it is the RP coefficients that we use to describe resulting values of reliability and to simulate policy scenarios.

19. Small, Winston, and Yan (2005b) discuss the procedures used to estimate these measures and to validate their accuracy.

A number of specification issues arise when we combine the RP and SP data sets. The RP analysis is described by equations (1), (2), and (3), to which we append superscript *RP* to distinguish those observations. The SP analysis, however, is different because we asked each respondent to express only a choice between the express or regular lanes, and we asked for this choice in eight different scenarios (each with different hypothetical values of travel variables). Because the SP choice is binary, it is convenient in the case of SP respondents to replace equation (1) by the utility *difference* between the express and regular lane. Thus in each choice scenario *t*, the respondent *n* chooses the express lane if, and only if

$$U_{nt}^{SP} \equiv X_{nt}^{SP}\beta_n^{SP} + \varepsilon_{nt}^{SP} \geq 0, \tag{6}$$

with β_n^{SP} given by equation (2) with the addition of *SP* superscripts on each of the symbols there. Note that β_n^{SP}, representing the preferences of individual *n,* does not vary across the different choice scenarios *t* presented to that individual.

We account for three additional effects that may arise due to the nature of the SP sample and to combining it with the RP samples. First, we expect the random influence ε_{nt}^{SP} to exhibit a typical panel structure. That is, it contains one random term, which we denote ξ_n, common to all the choice scenarios considered by individual *n*; and another, denoted η_{nt}^{SP}, that is unique to each choice scenario. Second, in the fifty-four cases where the same individual answered both the RP and SP questions, we expect some correspondence between the unobserved influences on their actual behavior and their hypothetical responses. To capture this, we assume that part of the random utility is common between them. Specifically, we assume the SP error ε_{nt}^{SP}, expressing random preference for the express lane as revealed in SP responses, contains a term proportional to ν_n^X from equation (3), representing random preference for travel in the express lane as revealed in observed (RP) behavior. Accounting for both of these effects results in the SP error in equation (6) given by:

$$\varepsilon_{nt}^{SP} = \xi_n + \theta\nu_n^X + \eta_{nt}^{SP}, \tag{7}$$

where

ξ_n: normally distributed with zero mean and variance normalized to one;
θ: a parameter to be estimated;
and η_{nt}^{SP} has a logistic distribution with standard deviation σ^{SP}.[20]

20. The logistic distribution describes the difference between two extreme-value variates, which is what we obtain because utility in equation (6) is the difference between the random utilities of express and regular lanes. Train (2003, p. 39) illustrates this point.

Finally, we follow standard practice in combining RP and SP data by allowing for differences between revealed and stated choices in the variance of random preferences. That is, the random factors affecting revealed choices may be larger or smaller than those affecting stated choices. For similar reasons, we allow for a difference between the two RP data sets in the variance of the random term in equation (1). The two data sets are Brookings RP (denoted BR) and CalPoly (denoted C). Thus the three standard deviations describing the three parts of our data (BR, C, and SP) are connected by two ratios that we estimate: $\tau^{BR} \equiv \sigma^{SP}/\sigma^{BR}$ and $\tau^{C} \equiv \sigma^{SP}/\sigma^{C}$. The ratios are described in our estimation results as *scale parameters.*[21]

We compute the log-likelihood function for our sample as the summation of logarithms of choice probabilities for RP observations (choice among nine alternatives) and for SP observations (binary choice), with the common error term ν_n^X entering both RP and SP choices for those people who are members of both Brookings samples. As noted, each choice probability is expressed in the usual manner for mixed logit as an integral of a multinomial or binary logit probability, conditional on normal random variates, over the distribution of those variates. We obtain parameter estimates by maximizing this log-likelihood function using Monte Carlo simulation to compute the integrals.[22]

Identification

Every statistical model must make explicit or implicit identifying assumptions about which environmental factors are held constant as others are varied, thus enabling the analyst to isolate the parameters of interest. Our model's parameters are identified by assuming that the unobserved influences on transponder, vehicle occupancy, and lane choice do not vary systematically by time of day. If they did, they would be correlated with the cost, time, and reliability of travel and their presence would bias those coefficients. The validity of this assumption depends to a large extent on how well our observed variables capture taste variation across time of day. Fortunately, it appears that such variation is reflected in several of our variables. For example, a motorist's gender is correlated with the time of day of travel. Females constitute only 15 percent of those people traveling during the interval 4:00 a.m.–

21. As in the binary logit model, one of these standard deviations can be normalized, typically by setting it equal to $\pi/\sqrt{3}$ for mathematical convenience: Train (2003), pp. 44–46. We normalize σ^{SP} in this way.

22. This is the maximum simulated likelihood estimator developed by Lee (1992) and McFadden and Train (2000), and exposited by Train (2003, pp. 148–9).

5:00 a.m., but 39 percent of the 7:00 a.m.–8:00 a.m. group. Similarly, the proportion of respondents whose trips are work trips varies from 100 percent at the earliest time to 58 percent at the latest time.

In earlier work, we conducted a formal test of whether unobserved taste variation by time of day affects our estimates of cost and travel-time coefficients.[23] The test consisted of estimating models of lane choice that included five time-of-day dummy variables. The findings indicated that values of time and reliability were not affected very much by the inclusion of the dummy variables. We do not include the time-of-day dummies in the current model because in the previous work they reduced the precision of the estimates.

Estimation Results

Table 3 presents estimation results. We group the RP parameters as those for generic variables that influence all three choice dimensions (transponder, lane, and vehicle occupancy), and those that influence just one of those dimensions. We also group separately those parameters influencing only the SP lane choice and those having a common effect on RP and SP choices.[24] Most influences are statistically significant and have the expected signs. As indicated by the generic RP coefficients and the SP coefficients, motorists pay close attention to the toll, travel time, and reliability when choosing among the available alternatives.

Observed heterogeneity in preferences is indicated by interaction variables formed by multiplying a generic variable by a socioeconomic or distance variable.[25] For example, toll is multiplied by a dummy variable for income, and median travel time is multiplied by various functions of trip distance. The results show that, as expected, motorists with higher incomes are less responsive to the toll, a statistically significant effect for RP respondents. The deterrent effect of travel time varies with distance in an inverted U-pattern,

23. Small, Winston, and Yan (2005a).

24. We conducted an extensive exploration of alternative specifications and functional forms for the explanatory variables, including removing the equality constraints between certain RP and SP parameters reflected in the "combined estimates" in the table. The model presented here is robust to such variations and is not rejected by statistical tests against more general models.

25. That is, we multiply a component of variable vector $\mathbf{X}_{jn}$ in equation (1) by a component of variable vector $\mathbf{W}_n$ in equation (2), as required by substituting equation (2) into equation (1). The resulting coefficient is a component of parameter vector $\boldsymbol{\gamma}$ in equation (2).

Table 3. Estimation Results for Demand Model

Variable	*Coefficient (standard error)*[a]
RP coefficients	
Generic variables	
Toll (dollars)[b, c]	–2.4042 (0.3994)
Toll[b, c] × dummy for high household income (> $60,000)	1.3869 (0.3395)
Median travel time (min.)[b] × trip distance (units of 10 miles)	–0.5753 (0.1751)
Median travel time[b] × trip distance squared	0.1128 (0.0394)
Median travel time[b] × trip distance cubed	–0.0050 (0.0020)
Travel-time uncertainty (80th percentile minus the median) (minutes)[b]	–0.7489 (0.2668)
Transponder choice	
Transponder dummy × Brookings dummy	–2.0101 (0.7472)
Transponder dummy × CalPoly dummy	–3.6342 (0.7374)
Female dummy × age 30–50 dummy × transponder dummy	1.8535 (0.7979)
Commute dummy × transponder dummy	1.2502 (0.6967)
Standard deviation of transponder dummy (σ^{T})	0.3276 (0.9422)
Lane choice	
Express lane dummy × Brookings dummy	0.2564 (1.1386)
Express lane dummy × CalPoly dummy	0.2264 (1.1691)
Standard deviation of express lane dummy ($\sigma^{X\text{-}RP}$)	3.7879 (0.8261)
Carpool choice	
Carpool dummy × Brookings dummy	–11.5192 (1.0339)
Carpool dummy × CalPoly dummy	–11.6719 (0.8883)
Female × age 30–50 × household size × carpool	1.4404 (0.3563)
HOV3 dummy × Brookings dummy	–9.2262 (0.9886)
HOV3 dummy × CalPoly dummy	–7.4263 (0.9909)
Common standard deviation of HOV dummies (σ^{HOV})	10.3225 (0.7837)
SP coefficients	
Express lane dummy	–3.0651 (1.1953)
Standard deviation of express lane dummy ($\sigma^{X\text{-}SP}$)	1.0530 (0.5237)
Toll (dollars)[b, c]	–1.4165 (0.3028)
Toll[b, c] × dummy for high household income (> $60,000)	–0.2492 (0.4808)
Travel time (minutes)[b] × long commute dummy (> 45 minutes)	–0.3538 (0.0812)
Travel time[b] × (1 – long commute dummy)	–0.3843 (0.0616)
Travel-time uncertainty[b]	–7.1139 (1.4507)
Combined coefficients	
Female dummy × express lane dummy	2.2434 (0.8384)
Age 30–50 dummy × express lane dummy	1.9277 (0.7955)
Household size (number of people) × express lane dummy	–0.7371 (0.2117)
Standard deviation of travel-time coefficient (σ^{Time})	0.3866 (0.0694)
Ratio of standard deviation to mean of coefficient of travel-time uncertainty ($\sigma^{Rel}/\gamma^{Rel}$)	1.3233 (0.3805)
Correlation parameter between RP and SP express lane choice (θ)	1.4808 (0.3209)
Parameters associated with scaling	
Scale parameter: CalPoly sample (τ^{C})	0.4143 (0.0902)
Scale parameter: Brookings RP sample (τ^{BR})	0.6064 (0.2029)

Table 3. Estimation Results for Demand Model (*Continued*)

Summary statistic	
Number of observations	1,124
Number of persons	538
Log-likelihood	1,059.63

Source: Based on data from the Brookings and CalPoly surveys and authors' calculations.

RP = Revealed preference data.

SP = Stated preference data.

a. Standard errors reported are the *sandwich* estimate of standard errors from Lee (1995). That is, each is the square root of the corresponding diagonal element in the matrix $\hat{V} = (-H)^{-1} P(-H)^{-1}$, where H is the Hessian of the simulated log-likelihood function and P is the outer product of its gradient vector (both calculated numerically). This estimate accounts for the simulation error in the likelihood function.

b. All cost, travel time, and unreliability variables are entered as the difference between values for the toll and free lanes. In the RP data, the cost for free lanes is zero, travel time for toll lanes is eight minutes, and unreliability for toll lanes is zero. In the SP data, cost, travel time, and unreliability are specified in the questions.

c. Value of cost for the toll lanes is the posted toll for a solo driver (for RP data) or the listed toll in the survey question (for SP), less 50 percent discount if car occupancy is three or more. For SP, car occupancy is determined from a question asking whether the respondent answered as a solo driver or as part of a carpool, and if the latter, what size carpool.

initially rising but then falling for trips greater than thirty-two miles. Following Calfee and Winston (1998), we conjecture that this pattern results from two opposing forces: the increasing scarcity of leisure time as commuting becomes longer and the self-selection of people with lower values of time into residences farther from work. For SP, we allow the coefficient on travel time to differ between people with long and short commutes, but we find the difference to be negligible.

We also find observed heterogeneity in alternative-specific preferences. The estimates listed under RP Coefficients in table 3 show that middle-aged females and all commuters are more inclined than other motorists to acquire a transponder. Middle-aged females with large families are more likely than other motorists to carpool, perhaps because they are more likely to make trips where family members ride together. Finally, as indicated by the combined coefficients, women, middle-aged motorists, and motorists in smaller households are more likely than others to choose the toll lanes, even after accounting for differences in transponder acquisition and car occupancy.[26]

Substantial unobserved heterogeneity in preferences over travel characteristics is indicated by the size and statistical significance of the estimated standard deviations σ^{Time} and σ^{Rel}. Similarly, there is unobserved heterogeneity

26. To better understand why women are more likely to use the toll lanes, we tried including an interaction of four variables: gender, age, household size, and either the express-lane dummy or the travel-time-uncertainty variable. This interaction sought to test whether working mothers with children might prefer the toll lanes or be more averse to unreliability due to tighter schedules. However, we could not find a measurable effect.

over absolute preferences for the express lanes ($\sigma^{X\text{-}RP}$) and for carpooling (σ^{HOV}). We tried also to estimate a random coefficient for the toll, but we were unable to obtain stable results. The standard deviations are estimated with good precision and are substantial in magnitude, ranging from roughly one-fourth of the corresponding mean coefficient to a multiple of it.[27] The scale and correlation parameters that describe the error structure are also estimated precisely and show, among other things, that the RP and SP responses from a single individual are strongly correlated.

We use our parameter estimates to compute properties of the distributions across individuals of motorists' implied value of time (VOT) and reliability (VOR).[28] In table 4 we provide summaries for all road users combined, and for users of the express lanes and free lanes separately. We use the Brookings RP sample for enumeration because it best represents the population, as argued above.[29] We characterize heterogeneity in VOT and VOR by the interquartile range (that is, the difference between the 75th and 25th percentile values) across individuals and across values of random parameters. This measure is relatively robust to the high upper tails typically found in distributions of ratios of random variables. The results are obtained by sampling across people in the enumeration sample and, for each, making repeated random draws from all estimated distributions.[30]

All of the 90 percent confidence intervals in the second column of table 4 are strictly positive, which indicates that all the reported estimates are statistically different from zero using a one-sided test at a 5 percent significance level. We find that the median value of time is \$19.63 an hour, which is about 85 percent of the average wage rate and thus near the upper end of the range expected from previous work.[31] The median value of reliability is \$20.76 an

27. The ratio of standard deviation to the mean coefficient is directly estimated for (un)reliability at 1.32. In the case of travel time, the estimated standard deviation of 0.39 may be compared with the SP coefficient of travel time of about −0.36 and with the derivative of utility with respect to RP median travel time, which is −0.76 at the mean trip distance of the Brookings RP sample.

28. These computations use the individual estimates for RP responses, which are derived according to equation (3) from the estimates of the mean coefficients from the section *RP Coefficients: Generic Variables* in table 3 and from the estimates of standard deviations from the section *Combined Coefficients.* Note that the latter estimates make use of both RP and SP responses.

29. There is no bias from choice-based sampling here because we use information only about respondents' characteristics, not their choices.

30. The same method is used, and described in greater detail, by Small, Winston, and Yan (2005a).

31. Small (1992, pp. 43–5).

Table 4. Implied Values of Time and Reliability for the Brookings RP Sample

Item	*Median estimate*	*90 percent confidence interval [5th, 95th percentile]*
Value of time (dollars/hour)[a]		
Median for:		
Entire sample	19.63	[8.75, 34.61]
Express lane users	25.51	[11.50, 39.99]
Free lane users	18.63	[7.76, 29.08]
Total heterogeneity[b] for:		
Entire sample	19.02	[12.57, 30.96]
Express lane users	29.30	[14.65, 55.97]
Free lane users	17.73	[11.37, 28.05]
Value of reliability (dollars/hour)[a]		
Median for:		
Entire sample	20.76	[8.37, 40.71]
Express lane users	23.78	[10.01, 48.29]
Free lane users	19.50	[5.73, 34.54]
Total heterogeneity[b] for:		
Entire sample	35.51	[14.95, 66.71]
Express lane users	44.70	[18.27, 84.24]
Free lane users	32.95	[13.70, 62.01]

Source: Based on data from the Brookings survey and authors' calculations.
RP = Revealed preference data.
a. Calculated from equations (4) and (5) using estimates of β_n^{Time}, β_n^{Rel}, and β_n^{Cost} applicable to RP responses (but note those estimates rely on both RP and SP data because they depend on σ^{Time} and σ^{Rel}).
b. Heterogeneity is expressed here as the interquartile range of the quantity in question across both individuals in the enumeration sample and random draws from the estimated distributions of the β's, in order to account for observed and unobserved heterogeneity, respectively.

hour.[32] Motorists also exhibit a wide range of preferences for speedy and reliable travel, as total heterogeneity in VOT and VOR is nearly equal to, or greater than, the corresponding median value. On average, express-lane users have higher values of travel time and reliability than do users of the free lanes, as expected, but wide and overlapping ranges exist within these two groups, resulting from the strong heterogeneity in preferences.

Simulating Highway Policies

We explore the policy implications of preference heterogeneity by developing a simulation model that uses our econometric results. It allows us to

32. Note that reliability is measured in hours because it is formed from percentiles of the distribution of travel time, which is measured in hours. Nevertheless, reliability, like travel time, is a property of the entire trip (more precisely, of the portion of the trip occurring on the section of the corridor we study).

examine current HOT and HOV policies and alternative pricing policies. We begin with a situation closely resembling the SR91 road-pricing experiment. Two ten-mile roadways, express and regular, are assumed to connect the same origin and destination and to have the same free-flow travel time of 8.0 minutes.[33] We model a four-hour peak period of 5:00 a.m.–9:00 a.m.

We find equilibria by iterating between the supply and demand sides of the model. The supply side is a standard static congestion model in which travel delays are proportional to the fourth power of the volume-capacity ratio.[34] Capacity is 2,000 vehicles an hour in each lane. Unreliability is assumed to be a constant fraction 0.3785 of travel delay (travel time minus free-flow travel time)—the fraction observed in our data on the free lanes averaged over 5:00 a.m.–9:00 a.m.

The demand side is obtained from the estimated demand model by sample enumeration, using the Brookings RP sample, which, as noted, is random and mostly representative of the population. The enumeration sample is assumed to represent a fixed population of size *N* potential commuters.

Our estimated demand model, of course, is conditional on travel in this corridor. However, we want to include the possibility of individuals altering their decision to travel in the corridor in response to policies we simulate, because other studies have shown that such responses can strongly affect the relative benefits of alternative pricing strategies.[35] Therefore we extend our estimated model by adding an outside choice representing nontravel (on the corridor).

The full procedure for this extension is described in the paper's appendix. Briefly, we postulate an outside or nontravel alternative labeled -1 (which could represent no trip, a trip outside the four-hour time period, or a trip on one of the other corridors that are some distance from the one we are modeling). Its utility is simply a constant $\overline{\delta}_{-1}$ plus a random term $\eta_{-1,n}$. The random terms for the other alternatives, that is, the terms η_{jn} in equation (3), are assumed to be more closely correlated with each other than with η_{-1}, just as in a nested logit model. A new parameter λ indicates the strength of this correlation. Choice probabilities (conditional on random parameters) are nested logit.[36]

33. This is the observed travel time on the section we study at 4:00 a.m., corresponding to a speed of seventy-five miles an hour.

34. U.S. Bureau of Public Roads (1964).

35. Verhoef, Nijkamp, and Rietveld (1996).

36. The choice probabilities are expressed in the appendix equations (A-2a-d).

Table 5. Initial Calibration of Expanded Demand Model

	Summer 2000 conditions		*More congested conditions*
Item	*HOT-lane policy*	*No-toll policy*	*No-toll policy*
Assumed toll[a]:			
Express lanes (dollars)	3.30	0	0
Free lanes (dollars)	0	0	0
Calibrated parameters:			
N	17,570	17,570	24,710
$\overline{\delta}_{-1}$	–12.65	–12.65	–23.41
Travel time (minutes):			
Express lanes	9.83	12.03	20
Free lanes	13.23	12.03	20
Arc elasticity of total corridor traffic volume with respect to full cost[b]	–0.40	–0.36	–0.36

Source: Authors' calculations.
a. HOV3+ pays half of the toll.
b. Based on 10 percent increase in time, unreliability, and cost for each travel option.

Calibrating the Expanded Demand Model

To use our model for simulation, we need to calibrate three parameters: the alternative-specific constant for the outside choice ($\overline{\delta}_{-1}$), the coefficient of the inclusive value of travel (λ), and the population size (N). Because we expect the travel alternatives to be much closer substitutes for each other than for nontravel, we choose λ as small as possible without causing numerical instability: namely, $\lambda = 0.2$. This choice does not seem to have much effect on the nature of the results. We calibrate the other two parameters ($\overline{\delta}_{-1}$ and N) to replicate observed traffic conditions during the morning peak on SR91 in the summer of 2000, which took place with an express-lane toll of \$3.30 with 50 percent discount for HOV3. The key traffic conditions are a travel-time difference between the express and free lanes of 3.4 minutes (according to our field measurements), and nontravel share of 10 percent.[37] The parameters that achieve these results are shown in the first column of numbers in table 5, along with the resulting travel times and the implied elas-

37. The 10 percent figure is a plausible estimate based on the likelihood that a small portion of trips were forgone due to congestion, that an alternative route available for some travelers (SR241) had about an 8 to 9 percent share of the CalPoly sample, and that the share of public transit is less than 1 percent.

ticity of traffic volume with respect to the full cost of travel.[38] The middle column shows the travel times produced by simulating the base (no-toll) policy with those parameters: namely 12.03 minutes, indicating a speed of fifty miles an hour.

We recognize, however, that most areas considering new pricing or express-lane policies have far worse congestion than was observed on SR91 in 2000—which was only five years after a 50 percent increase in the road's capacity. We therefore raise population N enough to reduce average speed to thirty miles an hour under a no-toll scenario. In setting the parameters for this starting situation, we hold constant not $\overline{\delta}_{-1}$, but instead the total traffic elasticity under a no-toll situation, shown in the last row of table 5. That elasticity (–0.36) may be compared with the value of –0.58 estimated under the actual pricing scheme in effect on SR91 in 2000, based on the CalPoly data and using a model with no heterogeneity.[39]

The calibration exercise just described leads finally to the parameters shown in the last column of the table, which we use in our policy simulations. We perform sensitivity tests, described later, using different values of the elasticity of total traffic volume, including a value of zero.

Defining Policies

Based on our equilibrium model of supply and demand, we simulate results for several pricing and operational policies. For each, we calculate tolls, travel times, traffic volumes, revenue, changes in consumers' surplus, and total change in social welfare. In our base-case (or no-toll policy), the two roadways are not distinguished. We compare policy scenarios that have the same number of total lanes (six), and thus do not investigate whether the benefits of a particular policy would merit adding new lanes in order to implement it.

38. Full cost is toll plus the traveler's value of travel time and unreliability faced. We computed the full-cost elasticity by using our expanded demand model, with initial calibrated parameters just described, to simulate changes in total travel under a no-toll scenario and a scenario where travel time and reliability are both increased 10 percent for all alternatives. We also computed the implied money-price elasticity of express-lane travel, which is −1.12, for comparison to the value of the same quantity reported in Yan, Small, and Sullivan (2002), which is –0.7 to –0.8 but based on a model that did not account for preference heterogeneity. We believe the higher elasticity calculated here is realistic because preference heterogeneity creates a subset of people who are quite ready to shift into or out of the express lanes in response to tolls, even though they are not very likely to shift from travel to nontravel.

39. Yan, Small, and Sullivan (2002).

The change in consumer surplus for traveler *n,* relative to the base case, is determined by the log-sum rule for nested logit:[40]

$$\Delta CS_n = \frac{1}{m_n} \Delta \ln\left[\exp\left(\bar{\delta}_{-1}\right) + \exp\left(\lambda I_n\right)\right], \quad (8)$$

where

Δ: indicates the difference between a given scenario and the base scenario;
I_n: inclusive value of the nine travel alternatives,
and m_n: individual's marginal utility of income, determined from the coefficient of the toll variable using Roy's identity.[41] The change in social welfare is the sum of expected changes in all individuals' consumer surplus and in toll revenues.

Besides the base policy (no toll), we first consider five other policies:

—HOV. A conventional carpool-lane policy in which the express lanes are open at no charge to carpools of two or more people.

—HOT. The express lanes are open both to carpools and to anyone willing to pay a toll.

—One-route toll. The express lanes are open to anyone willing to pay a toll, but with no discount for carpools.

—Two-route toll. All lanes are tolled, but with two different toll levels.

—Two-route HOT. Same as two-route toll except carpools can use either type of lane without charge.

For those policies requiring a toll, the toll is chosen to maximize social welfare subject to the constraints that define the policy. In the case of the HOT lane, the resulting optimal toll is smaller than would be charged under criteria typically used in current implementations.

The specific alternatives that are available for each policy are enumerated in table 6. Because their availability varies across policies, we must consider a feature of welfare analysis using discrete choice models, namely, adding more options increases welfare beyond any improvement in travel conditions that may actually result. This feature derives from the nature of a random utility model that assumes there are unobservable characteristics, captured

40. Choi and Moon (1997).

41. The equation for I_n is given as appendix equation (A-2c). We report all results on a per-trip basis, so Roy's identity equates m_n to minus the coefficient of the toll. Based on the results of table 3, $m_n = -(-2.4042+1.3869*H_n)$, where H_n represents a traveler's value of the high household income dummy.

Table 6. Availability of Alternatives by Policy[a]

	Alternative			*Policy*			
	Description						
Number	*Mode*	*Transponder?*	*Lane*	*No toll*	*HOV*	*HOT; one-route toll*	*Two-route toll; two-route HOT*
0	Solo	N	G	x	x	X	
1	Solo	T	G			X	x
2	Solo	T	X			X	x
3	HOV2	N	G	x	x	X	
4	HOV2	T	G			X	x
5a	HOV2	N	X		x		
5b	HOV2	T	X			X	x
6	HOV3	N	G	x	x	X	
7a	HOV3	T	G			X	x
7b	HOV3	N	X		x		
8	HOV3	T	X			X	x

Source: Authors' descriptions.

a. An x means that the alternative is available in that scenario; T indicates a transponder; N indicates none; G indicates the general lanes; X indicates the express lanes.

by ε_{jn} in equation (1), that differ among alternatives. We see from table 6 that some of our policies offer options with or without a transponder, while others offer express lanes as well as regular lanes. Our demand model was estimated in a situation including all nine possible alternatives described earlier, but when we simulate other policies, certain of these alternatives are eliminated from the choice set. This affects expected utility because the unobservable characteristics of eliminated alternatives are valued by some travelers. For example, some travelers in a HOT-lane environment value the day-to-day flexibility of lane choice that owning a transponder provides, whereas they do not have such flexibility in the no-toll or HOV policies.

Simulation Results

Table 7 shows simulation results. To facilitate understanding of the findings, we begin by presenting a detailed summary of the HOT-lane policy. The welfare-maximizing express toll for this policy is \$9.23 a trip (first row). It produces a big reduction of travel times on the express lanes compared with the base policy (from 20.0 to 12.4 minutes), and a much smaller reduction on the general lanes (from 20.0 to 19.2 minutes). The next set of numbers gives the shares of selected combinations of alternatives (in which the shares for alter-

Table 7. Policy Simulation Results

Effect	*No toll*	*HOV*	*HOT*	*One-route toll*	*Two-route toll*	*Two-route HOT*	*Limited two-route HOT*
Toll on express lane (dollars)	0	0	9.23	8.69	10.14	6.33	9.65
Toll on general lane (dollars)	0	0	0	0	8.16	5.34	1.90
Travel times (minutes):							
Express lane	20.00	11.8	12.4	11.6	11.6	13.1	12.5
General lane	20.00	18.8	19.2	22.6	12.8	12.5	16.5
Aggregated choice shares (percent):							
No travel on corridor	7.4	3.3	3.4	6.4	8.5	3.0	3.3
Solo on express lane	24.8	0	2.6	8.9	8.0	7.6	1.7
Solo on general lane	49.6	52.0	52.5	54.5	27.9	25.0	46.3
HOV2 on express lane	5.1	32.6	30.0	15.4	16.3	23.1	31.7
HOV2 on general lane[a]	10.3	2.9	3.0	7.2	22.1	26.8	6.6
HOV3+ on express lane	0.9	8.8	8.1	6.7	8.5	6.9	8.9
HOV3+ on general lane[b]	1.9	0.6	0.4	0.9	8.8	7.6	1.4
All HOV3+	2.8	9.4	8.5	7.6	17.3	14.5	10.3
All HOV	18.2	44.8	41.5	30.2	55.7	64.5	48.6
Consumer surplus change (dollars/person)[c]:							
Average	0	2.11	2.01	0.50	−2.36	0.98	1.36
Distribution in population (percentile)							
75th	0	2.92	2.71	0.65	0.00	3.51	2.80
50th	0	0.77	0.62	−0.27	−2.68	0.52	0.33
25th	0	0.26	0.26	−0.98	−5.36	−1.91	−0.98
Toll revenue (dollars/person)[c]	0	0	0.24	1.64	5.35	1.81	1.05
Welfare change (dollars/person)[c]	0	2.11	2.25	2.14	2.99	2.79	2.41

Source: Authors' calculations.

a. In the HOT and one-route toll policies, this row combines shares for two alternatives, with and without a transponder: namely, alternatives 3 and 4 in table 6.

b. Same as note a, but combines alternatives 6 and 7a in table 6.

c. Consumer surplus and social welfare are measured relative to the no-toll scenario. These two and toll revenue are each divided by the total number of potential travelers *N*. Social welfare is the sum of consumer surplus plus revenue.

natives with and without a transponder are added together). Thus for example, 3.4 percent of the *N* potential travelers choose not to travel on this corridor; just 2.6 percent pay the toll in order to travel alone on the express lanes (alternative *TX*1); and 52.5 percent travel alone in the general lanes (alternatives *TG*1 and *NG*1).

The consumer-surplus change for this policy, relative to No toll, averages \$2.01 a person—largely reflecting the reduced congestion caused by the shift to carpools (which is nearly as great as in the HOV policy, given that solo vehicles must pay a high express toll). The average, however, masks a wide dispersion in the gains to travelers, indicated by the percentiles shown next: the median traveler gets a 62-cent increase in consumer surplus, whereas the 75th percentile traveler gets a much larger \$2.71 while the 25th percentile traveler gets only 26 cents. Finally, the toll revenue collected from the HOT lane is just 24 cents a person, reflecting the small percentage paying the high toll. Adding the average consumer surplus change of \$2.01 to the average toll revenue of 24 cents gives the total welfare change per person, \$2.25, shown in the last row of table 7.

We now compare the welfare properties of the policies. The introduction of HOV lanes improves efficiency by encouraging carpooling—more than doubling its share from the base case (table 7 row labeled *All HOV*). This policy significantly reduces travel time on the express lanes, but leaves the general lanes very congested (table 7, third and fourth rows). In all likelihood, the policy would be much less effective in a smaller metropolitan area. Dahlgren (1998) finds that HOV lanes are favorable (in terms of reducing total person delay) only when initial congestion is substantial (delays of thirty-five minutes or more) and when the initial modal share of carpools is sizable (20 percent or more). We explore the welfare properties of HOV lanes under these conditions when we perform sensitivity analysis.

Allowing solo motorists to use the express lane if they pay a toll (HOT) generates a small welfare improvement over the HOV policy by enabling a small share of travelers to switch lanes and drive faster. In the HOT-like policy without an HOV exemption (one-route toll), the general lane becomes even more congested and the welfare improvement over the initial HOV policy is negligible. The two-route toll and two-route HOT policies generate considerably more welfare gains, as expected. But they do so at the cost of imposing large consumer-surplus losses on many travelers: the distribution in the population shows that the two-route toll reduces consumer surplus for three-fourths of all travelers (because the 75th percentile traveler gains \$0.00), while the two-route HOT reduces it for between one-fourth and one-half of them (because the 25th percentile gain is negative but the median gain is positive).

Table 8 provides more perspective on the policies' distributional effects by showing how consumer surplus varies between and within two different income groups. The dispersion within each group is quite large. For exam-

Table 8. Consumer Surplus Distribution by Income Group

Dollars per person	*No toll*	*HOV*	*HOT*	*One-route toll*	*Two-route toll*	*Two-route HOT*	*Limited two-route HOT*
High income (≥ $60,000) (percentile)							
75th	0	6.47	6.80	6.01	5.16	8.30	6.88
50th	0	1.68	1.61	1.27	0.92	4.48	2.47
25th	0	0.72	0.67	–0.92	–3.34	0.15	0.04
Low income (< $60,000) (percentile)							
75th	0	2.50	2.10	0.29	–0.55	2.47	2.00
50th	0	0.64	0.51	–0.37	–3.20	0.11	0.08
25th	0	0.22	0.22	–0.98	–5.60	–2.19	–1.04

Source: Authors' calculations.

ple, in the HOT policy the interquartile range is from 67 cents to $6.80 for the high-income group and from 22 cents to $2.10 for the low-income group. Note that these distributions show a great deal of overlap. For example, the one-quarter of the low-income group with the largest gains (that is, those above the 75th percentile) receive consumer surplus benefits of $2.10 or more, exceeding the gains of those at or below the median in the high-income group. As indicated previously, such findings reflect the considerable heterogeneity in VOT, VOR, and alternative-specific preferences that we found even controlling for income. Evidently there are many reasons besides income why some travelers strongly prefer one option over another.

Notwithstanding their sacrifice of economic efficiency, variants of HOV, HOT, and one-route pricing policies have attained a certain degree of public acceptance, suggesting that their distributional features are compelling enough to allow implementation. The first-best policy of tolling both lanes (two-route toll) produces a sizable gain in welfare over HOV and HOT policies, largely because it greatly reduces congestion on both lanes. However, it causes travelers to suffer high and disparate losses in consumer surplus, averaging $2.36 per trip and over $5.36 per trip for one-fourth of all travelers. Furthermore, the largest losses are associated with the lowest income groups, who tend to have the lowest values of time and reliability. These features obviously contribute to efficient pricing's lack of political appeal.

However, we find that policymakers can achieve most of the gains from first-best pricing, while partly addressing distributional concerns, by adding a carpool exemption to the two-route pricing policy (making it two-route

HOT).[42] The carpool share, already high in two-route pricing because of its financial incentives, increases even more, while congestion on both routes compares with the levels under two-route pricing. Remarkably, travelers, on average, obtain a substantial gain in consumer surplus (98 cents) from two-route HOT (compared with no toll); nevertheless, the policy is vulnerable to the concern that a substantial fraction of people incur sizable losses—those in the most disadvantaged quartile of users lose at least $1.91 per trip.

It is important to point out that if we assumed that travelers were homogeneous, our findings would change considerably, along the same lines as in other studies.[43] With homogeneous preferences, the relative welfare gain from the one-route toll, whose efficiency relies mainly on creating differential services, would drop significantly; HOT would become nearly identical in effect to HOV (because no additional benefits would result from separating users based on their preferences); and the one-route toll would be set much lower (because it cannot rely on attracting users just from the upper tail of the VOT and VOR distributions). In general, the reason that accounting for heterogeneity greatly affects policy comparisons is that diversity in users' preferences creates the opportunity to improve social welfare by providing differentiated services.

Toward a Better Policy Compromise

Our findings on the distribution of benefits and costs raise the question of whether it is possible to craft a policy that achieves an even better compromise between efficiency and political feasibility than the policies explored thus far. In particular, we seek a more efficient policy with the same attractive distributional features as the one-route toll. We choose one-route toll as a benchmark for political feasibility because at least two cases exist in North America where a policy resembling it has been successfully implemented. One case is SR91 itself. Although carpools did not all pay full fare on SR91, those with only two occupants did and those with three or more occupants paid half the fare during most of the time when the original private toll road

42. This policy, like others involving an express lane, incorporates the absolute preferences for the express lane from our demand model, which on average are slightly positive. For this reason, traffic is equilibrated when the express lane is slightly slower than the general lane, even though the express lane has a (moderately) higher price. While this may appear anomalous, it reflects other advantages of the express lanes on SR91, such as lack of trucks and intermediate entrances and exits, which we think would apply in many other express-lane applications.

43. Small and Yan (2001); Verhoef and Small (2004).

was in operation.[44] The other case is Highway 407 in the Toronto area. This is a publicly built highway (later sold to a private firm) that runs through the suburbs paralleling (a few miles away) a very congested east-west route through the city known as Queen Elizabeth Way. Projects such as these, and several others being actively considered, suggest that the public is willing to tolerate a toll road without HOV exceptions if a free alternative is available. In our simulation of the one-route toll, the free alternative exists in the form of a roadway immediately adjacent to the priced one, so this policy should be at least as acceptable as Highway 407.

We therefore quantify a benchmark for political viability as the 25th percentile of consumer-surplus change experienced by travelers under the one-route toll policy. Table 7 shows that this is –98 cents a trip. We then define an alternate version of two-route pricing that sets the two toll levels to maximize welfare, subject to the consumer surplus loss of the 25th percentile traveler not exceeding 98 cents a trip. The result is the limited two-route HOT policy shown in the last column in tables 7 and 8. It results in a sharply differentiated toll: for the express lane its magnitude ($9.65) compares with that in the two-route toll policy, but for the general lane it is only $1.90, much lower than either of the other two-route pricing policies. The policy achieves a general-lane speed of thirty-six miles per hour, intermediate between those of the no-toll and two-route policies. It also achieves greater efficiency than any of the policies that leave the general lanes unpriced.

Motorists in the limited two-route HOT policy achieve a consumer surplus gain, on average, of $1.36 compared with no tolls. This exceeds the gain achieved by any other two-route pricing policy. Furthermore, travelers are treated much more evenly than in the other two-route pricing policies, with the interquartile range only modestly greater than with HOT or HOV. Thus the limited two-route HOT policy succeeds in improving efficiency more than most other policies, while maintaining the attractive distributional characteristics comparable to policies that have been found to be politically feasible.

We stress that our policy simulations are based on an experiment concerning only a single ten-mile stretch of highway. Most significant congestion affects a

44. Although there were complaints about the high tolls and about charging HOV3+ vehicles, they do not appear to have undermined the stability of the arrangement, which had a strikingly high acceptance level in various polls. What did eventually undermine the private operation was an unrelated issue: the franchise allowed the private operator to veto any capacity improvements in the corridor, which it did through a lawsuit in a highly publicized dispute with the California Department of Transportation. As a result, the express-lane franchise was purchased in 2003 by a public agency, which has retained most of the toll policies of its private predecessor.

much broader region. If the distributional advantages of differentiated pricing enable it to be broadly adopted, its welfare gains will be greatly magnified.

Sensitivity Analysis for Simulation Results

The simulation results presented so far are based on a full-cost elasticity for total traffic of –0.36. Appendix tables A-2 and A-3 present simulation results where we assume the full-cost elasticity of total corridor traffic is –0.60 and zero, respectively. For the case of zero elasticity, the model has no outside option and so there is no parameter $\overline{\delta}_{-1}$ to calibrate. In the other case, the two parameters N and $\overline{\delta}_{-1}$ are simultaneously calibrated, as in the main results, to achieve the desired elasticity and travel-time differential with no toll. The results show that the welfare rankings of various policies, and the nature of their distributional impacts, do not depend on this assumed elasticity. We caution that specific numerical results are not necessarily comparable with different assumed elasticities because they imply different starting shares for HOV.

A more important area for sensitivity analysis, in our view, is varying the initial carpool share. As noted, many metropolitan areas have much smaller carpool shares than Los Angeles, a large city that has relatively long work trips. We therefore present an alternate simulation in which we change the parameters governing the alternative-specific utilities for HOV alternatives so as to produce HOV shares about half those of our primary scenario (table 7), under both the no-toll and HOV policies.[45] Results are shown in table 9. The policies do not perform as well as those in our main simulation, either in total welfare gain or in direct impact on consumers. For example, the median consumer-surplus change is uniformly negative. The smaller efficiency gains and less favorable distributional properties occur because the policies cannot induce as much carpool formation and therefore they achieve less relief from congestion. The HOV policy, especially, is much less effective—it produces a negative total welfare change and substantial consumer surplus losses both to the median and, especially, the 25th percentile traveler. This striking finding is consistent with policymakers' growing dissatisfaction with HOV lanes and growing interest in HOT lanes throughout the country.

45. We do this by adjusting the HOV dummy and its standard deviation. (The HOV3 dummy is adjusted proportionally to the HOV dummy.) The HOV share can be estimated approximately for about twenty existing corridors with HOV lanes in U.S. metropolitan areas from data in Pratt and others (2000, tables 2-2, 2-7, and 2-9). Almost all are between 15 and 30 percent, whose midpoint (22.5 percent) is almost exactly half the share predicted by our HOV policy simulation in table 7.

Table 9. Simulation Results with Low HOV Share[a]

Effect	*No toll*	*HOV*	*HOT*	*One-route toll*	*Two-route toll*	*Two-route HOT*	*Limited two-route HOT*
Toll on express lane (dollars)	0	0	8.03	8.20	11.10	9.47	8.35
Toll on general lane (dollars)	0	0	0	0	7.94	7.02	1.98
Travel times (minutes):							
Express lane	20.00	9.31	12.38	12.28	12.07	13.37	12.64
General lane	20.00	25.38	22.95	24.04	15.87	15.50	20.75
Aggregated choice shares (percent):							
No travel on corridor	7.49	13.31	7.84	9.26	18.37	13.23	9.08
Solo on express lane	27.88	0	14.66	17.75	16.80	15.57	14.59
Solo on general lane	55.77	64.09	61.43	61.55	48.38	47.43	58.47
HOV2 on express lane	2.47	16.71	11.84	5.90	6.70	12.70	12.58
HOV2 on general lane	4.94	1.12	1.21	3.22	5.89	6.26	1.72
HOV3 on express lane	0.48	4.61	2.77	1.85	2.41	3.29	3.17
HOV3 on general lane	0.97	0.16	0.25	0.47	1.45	1.52	0.39
All HOV3	1.45	4.77	3.02	2.32	3.86	4.81	3.56
All HOV	8.86	22.60	16.07	11.44	16.45	23.77	17.86
Consumer surplus change (dollars/person):							
Average	0	–0.50	0.41	–0.20	–4.22	–2.81	–0.67
Distribution in population (percentile)							
75th	0	0.30	0.68	0.28	–1.98	–0.16	0.00
50th	0	–1.12	–0.35	–0.69	–5.27	–3.86	–1.98
25th	0	–2.51	–1.14	–1.66	–7.01	–5.81	–2.51
Toll revenue (dollars/person)	0	0	1.18	1.75	6.44	4.82	2.38
Welfare change (dollars/person)	0	–0.50	1.59	1.55	2.22	2.01	1.71

Source: Authors' calculations.
a. See notes for table 7.

What about the limited two-route HOT policy? Can it outperform HOT? Although we found that it cannot meet its political acceptability, as indicated by HOT's 25th percentile consumer surplus loss, it can improve on HOT's efficiency if we define political acceptability relative to the 25th percentile consumer-surplus loss of HOV instead. Thus the possibility for an improved policy compromise exists for other metropolitan areas in the country, although we cannot be as sanguine about political feasibility as before.

Conclusion

Methodological advances in microeconometrics have enriched our understanding of consumer behavior by recognizing that consumers are not homogeneous. Applications have shown that accounting for heterogeneity is important when assessing policies in many domains, such as economic deregulation, job training, and poverty programs. We find that heterogeneity plays a similarly important role in policy toward highway transportation. Accounting for it creates the opportunity not only to introduce HOT lanes, as has been previously recognized, but to introduce more far-reaching pricing policies within the limits of distributional effects that appear to be politically acceptable in certain circumstances. We have been able to design a differentiated road-pricing scheme that fills in the gap between optimal but socially unpopular first-best pricing and pragmatic but less-efficient policies like carpool or HOT lanes.

Recent experiments have shown that policymakers are no longer unwilling to use the price mechanism to allocate scarce road capacity. The changing times give cause for optimism that more efficient policies, offering choices that appeal to diverse users, will become serious candidates for implementation.

APPENDIX

Stated Preference Survey Questionnaire

Eight hypothetical commuting scenarios were constructed for respondents who travel on California State Route 91. Respondents who indicated their actual commute was less (more) than forty-five minutes were given scenarios that involved trips ranging from twenty to forty (fifty to seventy) minutes. An illustrative scenario is presented in table A-1.

Extended Demand Model for Simulations

Let $\Omega = \{-1,0,1, \ldots, 8\}$ denote the choice set for a potential road user, where alternative -1 is the outside choice and alternatives 0–8 represent the

Table A-1. Stated Preference Survey Questionnaire

Free lanes	*Express lanes*
Usual travel time: Twenty-five minutes	Usual travel time: Fifteen minutes
Toll: None	Toll: $3.75
Frequency of unexpected delays of ten minutes or more: One day in five	Frequency of unexpected delays of ten minutes or more: One day in twenty
Your choice (check one): Free lanes ☐	 Toll lanes ☐

Source: Authors' descriptions.

different combinations of routes, transponder acquisition, and car occupancy defined previously. It is convenient to let $\tilde{\Omega} = \{0,1, \ldots, 8\}$ denote the subset of choices involving travel on the corridor.

The utility of individual n choosing alternative j is:

$$\text{(A-1a)} \qquad U_{-1n} = \bar{\delta}_{-1} + \eta_{-1n},$$

$$\text{(A-1b)} \qquad U_{nj} = X_j^{RP}\beta_n^{RP} + \varepsilon_{jn}^{RP}, j \geq 0,$$

with β_n^{RP} and ε_{jn}^{RP} as given by equations (2) and (3) in the text. Thus each traveler's utility for nontravel is divided into a mean $\bar{\delta}_{-1}$, which is constant for all commuters, and random deviation ε_{-1n}; whereas the utility for each alternative involving travel is the same as in the RP part of our estimated model, equations (1), (2), and (3) in the text. We henceforth omit the superscript *RP* for simplicity.

The random preferences for individual n therefore are represented by the vector $\boldsymbol{\Psi}_n = (\boldsymbol{\eta}_n, \boldsymbol{\nu}_n, \boldsymbol{\mu}_n)$, where $\boldsymbol{\eta}_n = (\eta_{-1n}, \eta_{0n}, \ldots, \eta_{8n})$, $\boldsymbol{\nu}_n = (\nu_n^T, \nu_n^X, \nu_n^{H2}, \nu_n^{H3})$, and $\boldsymbol{\mu}_n = (\mu_n^{Time}, \mu_n^{Rel})$. The density function of $\boldsymbol{\Psi}_n$ is specified as $\rho(\boldsymbol{\Psi}_n) = \rho_\eta(\boldsymbol{\eta}_n) \cdot \rho_{\nu\mu}(\boldsymbol{\nu}_n, \boldsymbol{\mu}_n)$. Here $\rho_{\nu\mu}(\cdot)$ is a product of independent normal random variables with standard deviations as estimated, while $\rho_n(\cdot)$ takes the nested-logit form in which the outside alternative -1 is one nest and the travel alternatives $\tilde{\Omega}$ are another nest with similarity parameter λ. This specification captures the idea that the substitution pattern between any two travel choices may be different from that between nontravel and travel. The market share of alternative $j \in \tilde{\Omega}$, within the submarket represented by people with characteristics of traveler n in our enumeration sample, is found by integrating the nested-logit probability formula, conditional on random parameters ν_n and μ_n, over the distribution function of those random parameters:

(A-2a) $$S_{jn} = \int_{(\nu_n,\mu_n)} S_{jn}^{(\nu_n,\mu_n)} \cdot \rho_{\nu\mu}\left(\boldsymbol{\nu}_n,\boldsymbol{\mu}_n\right) d\left(\boldsymbol{\nu}_n,\boldsymbol{\mu}_n\right),$$

where

(A-2b) $$S_{jn}^{(\nu_n,\mu_n)} = \frac{\exp\left(X_{jn}\beta_n/\lambda\right)}{\exp\left(I_n\right)} \cdot \frac{\exp\left(\lambda I_n\right)}{\exp\left(\bar{\delta}_{-1}\right) + \exp\left(\lambda I_n\right)}$$

is the share conditional on values of the normal random variates, and

(A-2c) $$I_n = \ln \sum_{j=0}^{8} \exp\left(X_{jn}\beta_n/\lambda\right)$$

is the inclusive value of travel choices. The nontravel share is

(A-2d) $$S_{-1n} = \frac{\exp\left(\bar{\delta}_{-1}\right)}{\exp\left(\bar{\delta}_{-1}\right) + \exp\left(\lambda I_n\right)}.$$

The total demand for an alternative j is therefore

(A-3) $$D_j = \sum_n w_n S_{jn},$$

where w_n is the number of people represented by motorist n. This number is just $w_n = N/79$, where N is the total population size, since our enumeration sample consists of seventy-nine equally weighted individuals. The traffic volume arising from those individuals who choose a travel alternative j involving occupancy O_j is $V_j \equiv D_j/O_j$.

Table A-2. Alternate Simulation Results: Elasticity = −0.60

Effect	*No toll*	*HOV*	*HOT*	*One-route toll*	*Two-route toll*	*Two-route HOT*
Toll on express lane (dollars)	0	0	8.41	8.53	9.41	6.02
Toll on general lane (dollars)	0	0	0	0	7.01	5.32
Travel times (minutes):						
Express lane	20.0	12.4	13.0	11.3	11.4	13.8
General lane	20.0	19.8	20.0	22.9	13.0	12.8
Aggregated choice shares (percent):						
Outside choice	16.5	9.8	10.8	15.9	19.7	9.8
Solo on express lane	22.2	0	2.8	7.4	7.2	7.9
Solo on general lane	44.3	47.9	48.0	48.9	26.7	21.9
HOV2 on express lane	4.8	30.8	27.9	14.1	14.1	20.4
HOV2 on general lane	9.6	2.7	3.0	6.9	18.8	26.7
HOV3 on express lane	0.9	8.5	6.9	5.9	7.2	6.0
HOV3 on general lane	1.7	0.4	0.6	0.9	6.3	7.3
All HOV3	2.6	8.9	7.5	6.8	13.5	13.3
All HOV	17.0	42.4	38.4	27.8	46.4	60.5
Consumer surplus change (dollars/person relative to No toll):						
Average	0	1.50	1.47	0.46	1.77	0.49
Distribution in population (percentile)						
75th	0	2.04	1.94	0.56	0.03	2.73
50th	0	0.10	0.03	–0.14	–1.76	0.08
25th	0	0.03	0.01	–0.87	–4.61	–1.94
Toll revenue (dollars/person)	0	0	0.24	1.40	4.24	1.63
Welfare change (dollars/person)	0	1.50	1.71	1.86	2.47	2.13
Consumer surplus distribution by income group (dollars/person)						
High income (≥ $60,000) (percentile)						
75th	0	5.04	5.04	5.16	4.85	6.67
50th	0	0.23	1.08	1.02	1.47	3.50
25th	0	0.08	0.02	–0.65	–2.28	0.18
Low income (< $60,000) (percentile)						
75th	0	1.75	1.42	0.20	–0.02	1.86
50th	0	0.09	0.02	–0.25	–2.26	0.00
25th	0	0.03	0.01	–0.89	–4.83	–2.28

Source: Authors' calculations.
Notes: See notes for table 7.

Table A-3. Alternate Simulation Results: Elasticity = 0 (No Outside Choice)

Effect	*No toll*	*HOV*	*HOT*	*One-route toll*	*Two-route toll*	*Two-route HOT*
Toll on express lane (dollars)	0	0	8.47	8.64	10.46	6.17
Toll on general lane (dollars)	0	0	0	0	9.29	5.19
Travel times (minutes):						
Express lane	20.0	11.0	11.6	11.6	11.8	12.3
General lane	20.0	18.1	18.3	22.1	12.4	12.3
Aggregated choice shares (percent):						
Outside choice	0	0	0	0	0	0
Solo on express lane	26.6	0	3.5	10.4	9.7	8.8
Solo on general lane	53.2	54.3	54.5	57.6	26.6	26.6
HOV2 on express lane	5.6	31.7	28.7	15.2	16.4	21.7
HOV2 on general lane	11.3	4.2	4.5	8.5	26.4	28.1
HOV3 on express lane	1.1	8.9	7.9	7.0	8.9	6.6
HOV3 on general lane	2.2	0.8	0.9	1.3	11.9	8.2
All HOV3	3.3	9.8	8.8	8.3	20.9	14.9
All HOV	20.2	45.7	42.0	32.0	63.4	64.6
Consumer surplus change ($/person relative to No toll):						
Average	0	2.94	3.09	1.17	–2.39	1.87
Distribution in population (percentile)						
75th	0	4.07	4.00	1.15	0.12	4.45
50th	0	1.47	1.64	0.14	–3.05	1.29
25th	0	0.56	0.91	–0.44	–5.78	–1.48
Toll revenue (dollars/person)	0	0	0.24	1.40	4.24	1.63
Welfare change (dollars/person)	0	1.50	1.71	1.86	2.47	2.13
Consumer surplus distribution by income group (dollars/person)						
High income (≥ $60,000) (percentile)						
75th	0	8.35	8.68	6.70	6.04	9.81
50th	0	3.06	4.02	1.79	1.49	5.30
25th	0	1.42	2.23	0.13	–3.10	0.93
Low income (< $60,000) (percentile)						
75th	0	3.38	3.04	0.71	–0.94	3.46
50th	0	1.22	1.40	0.01	–3.60	0.77
25th	0	0.49	0.82	–0.50	–6.08	–1.73

Source: Authors' calculations.
Notes: See notes for table 7.

Comments

Nathaniel Baum-Snow: Small, Winston, and Yan's paper provides valuable new insights on the potential social welfare gains associated with implementing various new urban highway lane pricing and carpooling options. The authors apply modern econometric techniques to estimate a partial equilibrium model of highway travel demand. Using the resulting estimates, the authors perform a detailed welfare analysis of different high-occupancy vehicle and toll lane policies. California State Route 91 (SR91) is an ideal case to evaluate. The road has few exits, considerable variation in congestion delays for different times of day, and variation across observations in the cost of the HOT lane. The structural approach lends itself well to the full welfare analysis of various policy alternatives performed at the paper's end. The authors choose their empirical specification to be flexible enough to inform us about distributional consequences of different HOV, toll lane, and HOT policies in addition to their mean effects. Finally, the authors deserve considerable credit for their data collection efforts. Without these unique data, this analysis would not have been possible. This paper should be of great use to urban transportation policymakers considering the implementation of highway congestion tolls and HOV lanes.

The authors cite a body of work documenting that road congestion delays have been rising rapidly since the early 1980s. At the same time, the fraction of the U.S. population commuting by car continues to increase, reaching 91 percent in 2000. Commuting by car has risen even in the face of large investments in public transit infrastructure. On the face of it, this may seem like a golden opportunity for carpool lanes. At very low cost, they have the potential to provide large congestion reductions, thereby increasing commuting speeds for many commuters. However, as seen in table 10, the fraction of auto commuters who carpool has been falling rapidly. In 1980 it stood at 20 percent, while by 2000 it had fallen to just 12 percent nationally. The decline in carpooling is seen for central city and suburban residents alike. It also holds

Table 10. Trends in Carpooling, 1980–2000[a]

Area			*1980*	*1990*	*2000*
United States	Central city	Fraction commute by car	0.75	0.75	0.62
		Car commuters who carpool	0.17	0.14	0.12
	Suburban	Fraction commute by car	0.90	0.92	0.92
		Car commuters who carpool	0.20	0.13	0.12
	All	Fraction commute by car	0.86	0.89	0.91
		Car commuters who carpool	0.20	0.14	0.12
Los Angeles CMSA		Fraction commute by car	0.88	0.89	0.90
		Car commuters who carpool	0.17	0.15	0.15

Source: Author's calculations using the 1980, 1990, and 2000 Census Public Use Microdata Samples (PUMS).
CMSA = Consolidated Metropolitan Statistical Area.
a. Sample includes all commuters in listed areas. Regions for which central city and suburban numbers are broken out vary over time.

in the Los Angeles Consolidated Metropolitan Statistical Area (CMSA), where the SR91 experiment took place, though to a lesser extent. In Los Angeles carpooling rates fell from 17 to 15 percent of drivers between 1980 and 2000. Over this period, the fraction of commuters using public transit declined by less in both absolute and percentage terms, despite considerable public investment in both transit and HOV lane infrastructure.

How can one explain steeply declining carpooling rates despite longer congestion delays and more drivers on the roads? A few clues are revealed by examining income levels and travel times. Table 11 presents average commuting times and total income levels (in 1999 dollars) of people who drive alone and people who carpool. It shows that carpoolers are considerably poorer than solo drivers and that their average commuting time is greater. This pattern holds in central cities and suburban areas, persists over time, and also holds for the Los Angeles CMSA. Two explanations may be important in accounting for these patterns. Poor people may have longer distance commutes and as a result may be more likely to find carpooling (and carpool lanes) to be useful. In addition, while carpooling reduces highway travel time where there are carpool lanes, it increases nonhighway travel time due to the need to circulate to pick up or drop off passengers. Population and employment decentralization can only have increased the fixed-time cost of carpooling. As such, the driving force behind more prevalent carpooling among the poor may be savings on pecuniary costs of car ownership and gasoline rather than time savings.

In the SR91 case, which involves longer than average commutes, about eight minutes is the maximum time savings available from carpooling under

Table 11. Income and Commuting Time of Solo Drivers and Carpoolers, 1980–2000[a]

	1980		*1990*		*2000*	
Area	*Solo driver*	*Carpooler*	*Solo driver*	*Carpooler*	*Solo driver*	*Carpooler*
United States						
Central cities						
Income	33,986	27,760	36,775	27,075	42,412	31,363
Minutes	20	24	21	25	26	30
Suburban area						
Income	36,551	31,925	40,129	31,133	47,830	33,740
Minutes	21	26	24	28	27	30
All						
Income	33,599	28,568	35,336	27,194	39,222	27,969
Minutes	20	25	21	26	23	27
Los Angeles CMSA						
Income	37,851	30,709	43,757	28,849	45,987	28,830
Minutes	23	27	25	28	27	29

Source: Author's calculations using the 1980, 1990, and 2000 Census Public Use Microdata Samples (PUMS).
CMSA = Consolidated Metropolitan Statistical Area.
a. Sample includes all those who reported commuting by car. Income reports total annual income adjusted to 1999 dollars, minutes are one-way commuting time.

any HOV or HOT lane scenario considered. Picking up and dropping off a carpooler, or even coordinating the departure of two people who live or work in the same location, can easily take this much time every trip. The suggestive evidence in tables 10 and 11 is consistent with the time cost of carpooling increasing faster as a function of the wage than the time cost of solo driving. This may reflect a large fixed-time cost associated with carpooling, implying that carpooling may only be attractive to the rich if they have very long commutes. Consistent with these observations, the authors estimate large negative marginal utilities of carpooling for the average SR91 commuter.

Given the bleak outlook for carpooling and dominant role of the car in commuting, it is natural to look for road pricing schemes that allow solo commuters to pay for the congestion externality they impose on others. Thus it is appropriate that the authors use a full congestion pricing scheme to estimate their model, and consider a broad set of policy alternatives.

The results presented in table 7 highlight a host of potential difficulties in selecting and implementing various toll policies. Since high-income people have the highest value of time and lowest value of money, they are the ones who will ultimately benefit the most from being able to pay for a faster commute. Full congestion pricing of the road leaves the poor with the costly options of paying a toll or using an alternate longer route, thereby making them worse off. With their lower value of time, the poor endeavor to trade off pecuniary costs for time. It is this same intuition that can explain much higher rates of bus usage among the poor than other groups. As the authors demonstrate in table 9, there has to be considerable flexibility to carpool in the population in order for a HOT lane scheme to generate welfare gains for all but those at the very top end of the income distribution. The HOT lane attracts those with a relatively low cost of carpooling who reduce congestion for everyone else, as well as solo drivers who have high values of time and reliability. In the authors' empirical model, those at the lower end of the income distribution only benefit from the congestion reduction due to carpooling; low-income commuters may be burdened with longer commutes if a HOT lane induces very few people to carpool as the free lane becomes more congested.

Given this friction between the interests of the rich and poor, what is the feasibility and efficacy of establishing HOT lanes or limited two-route HOT lanes in other areas? With current carpooling rates considerably less than the baseline of 18 percent in the SR91 case, one may find that implementing HOT lanes in other areas would have much more deleterious effects for the poor than seen in the SR91 case. Policymakers should thus be cautious in applying Small, Winston, and Yan's estimates to other situations. However, reestimation of the model with new stated preference data that include information on carpooling propensities for areas that are studying implementation of HOT lanes likely would be sufficient to perform the appropriate welfare analyses.

There are a few important considerations that fall beyond the scope of this paper. Congestion pricing of highways has the possibility of mitigating inefficient spatial dispersion in urban areas. Indeed, as noted by the authors, the negative externalities imposed by extra highway users not only manifest themselves as slower commute times for their commuting brethren, but also in inefficient land use patterns. Indeed, the general equilibrium consequences of improvements in the transportation infrastructure can be quite important, as seen in the influence highway construction had on suburbanization. The other general equilibrium response not considered by the authors but that may be important is commuting time-shifting. If poor individuals can adjust their commuting hours at low cost to avoid tolls, the authors may be underestimating the welfare benefits of road congestion tolls.

Overall, one learns a lot from Small, Winston, and Yan's work. It would be interesting to use their estimates, or new results using the same methodology estimated using data from a stated preference survey, to form predictions about lane choice and carpooling for projects coming on line in the next few years. Successful out-of-sample predictions for travel demand would lend further credence to the welfare analysis performed in this paper. With any luck, the methodology developed in this paper will be adopted by policymakers for evaluating potential road pricing schemes in the future.

José A. Gómez-Ibáñez: Small, Winston, and Yan set out to demonstrate, using traffic congestion as their example, that public policy can be improved by understanding how preferences vary among consumers. Local officials have long resisted the suggestion by many economists that the problem of urban traffic congestion can be significantly reduced by charging motorists tolls that depend on the level of congestion. Officials have been much more willing to build special high-occupancy vehicle (HOV) lanes to try to reduce congestion by encouraging the formation of carpools. In addition, some communities have converted their HOV lanes into high-occupancy toll (HOT) lanes, usually in response to criticism that the HOV lanes are underused while the general-purpose lanes remain very crowded. HOT lanes allow motorists who are driving alone (or who do not have enough passengers to qualify for HOV status) to use the exclusive lanes if they pay a toll. Only a few major urban areas (notably Singapore, in 1975, and London, in 2003) have been willing to toll all lanes, a practice that economists argue would establish strong incentives for motorists to form carpools, shift to mass transit, and travel in less-congested hours or to less-congested locations.

Small, Winston, and Yan cite research showing that HOV and HOT lanes can actually increase rather than reduce traffic congestion, particularly in circumstances where the lanes do not encourage the formation of many new carpools or the new carpools are formed primarily by former bus riders. The authors argue that HOV and HOT lanes are both inferior to congestion tolls on all lanes because the general-purpose lanes are left untolled and thus excessively congested. They contend that policymakers can fashion a politically acceptable policy that generates most of the benefits of congestion tolls by taking advantage of the variation in preferences for different commuting options. The basic idea is to offer a menu of options instead of a single policy, so that consumers can pick the option that reduces congestion at the least cost and inconvenience to them. The authors illustrate their argument by simulating alternative congestion-relief policies on a stretch of California State Route 91 (SR-91) that currently has four general-purpose and two HOV lanes.

Transportation planners and policymakers will not need much convincing about the importance of considering heterogeneous preferences. They have long understood, for example, that variations in preferences are the reason that several modes of transportation often serve the same territory or route. Trucks and railroads both can thrive on the same route because some shippers are more sensitive to travel time and reliability, while others are more sensitive to cost. Similarly, buses, trains, and automobiles often share the same commuter corridor because commuters differ in the value they place on travel time savings, their willingness to carpool, their need to run errands on the way to or from work, and other characteristics. Transportation planners commonly break down markets into different segments that vary not just by the origins and destinations of their trips but by their different preferences. Small, Winston, and Yan are encouraging planners to consider the implications of preference heterogeneity in a wider variety of applications.

Small, Winston, and Yan's simulations make it clear why economists have had such a difficult time convincing policymakers to impose congestion tolls. Congestion tolls on both the general-purpose and special lanes of SR-91 generate the largest estimated social welfare gain of any policy they simulate: \$2.99 per commuter. But the average commuter actually loses \$2.36 in consumer surplus from congestion tolling—only a handful of wealthy commuters who place a high value on time savings are better off. The reason that congestion pricing generates net benefits to society is that the government collects tolls averaging \$5.35 per commuter. Who will benefit from the toll revenues depends on how the revenues are used. If they are used to cut taxes, then taxpayers benefit, and if they are used to expand government programs, then program users benefit. But the fear is that the benefits from toll revenues are likely to be spread so thinly over a large group of taxpayers and government program users that they will be hardly noticed, while the costs will be concentrated on commuters and very visible. Diffuse benefits and concentrated costs are a recipe for political resistance.

The most obvious solution to this dilemma is to attempt to return some of the toll revenues to commuters in a way that does not undermine the incentive properties of the tolls. For example, London made its congestion tolling scheme more popular by pledging to use the toll revenues to expand bus services.[1] Similarly, Hong Kong promised to use congestion toll revenues to reduce auto-

1. A more important factor in public acceptance of the London scheme was probably that 85 percent of the commuters working in the central area, where congestion tolls were applied, were already using public transit.

mobile registration fees and excise taxes. Hong Kong's proposal failed, in part, because motorists did not trust the government to keep its promises to cut auto taxes.

This paper's authors suggest a new strategy: taking advantage of the variation in consumer preferences for time savings and other attributes to set congestion tolls at levels that benefit, rather than harm, most motorists. The fact that there are two sets of lanes (special and general-purpose) allows Small, Winston, and Yan to offer two toll levels simultaneously: a high toll in the special lanes that delivers a substantial time savings and appeals to motorists who value time greatly, and a low toll in the general-purpose lanes level that delivers less time savings but is more attractive to motorists who place less value on travel time. In addition, carpools can travel free in the high toll lane. The authors tinker with the two toll levels until they find a pair of tolls that gives a net social gain of $2.41 per commuter, or 81 percent of the $2.99 maximum benefit that might be obtained. But with their tolls the average commuter gains a consumer's surplus of $1.36 instead of losing $2.36. Only roughly a third of commuters lose, and they typically lose only $1 per trip.[2]

While the authors' solution is very clever, they do not recognize that the designers of conventional HOT lanes have essentially anticipated their idea. Conventional HOT lanes also offer commuters three different options: untolled but slow travel in the general-purpose lanes (which appeals to those who do not value travel time savings highly); untolled and fast travel in the special-purpose lanes (which appeals to those who do not mind traveling in carpools); and tolled and fast travel in the special-purpose lanes (for those who highly value time savings and prefer not to carpool). As a result, the HOT lane policy generates a net social gain of $2.25 per commuter, or 75 percent of the $2.99 maximum possible and almost as much as the $2.41 generated by the authors' scheme. Equally important, the average commuter gains $2.01 instead of losing $2.36, and virtually all commuters are winners.[3] In short, the HOT lane scheme uses the same approach of providing a menu of options as the authors' proposal and, as a result, gains almost the same net social benefit and an even better outcome for commuters.

Small, Winston, and Yan might have made a stronger case for their particular proposal had they used a more representative highway than SR-91. This California road is ideally suited for HOV and HOT lanes because the commuters that use it travel very long distances. SR-91 crosses a barrier of hills

2. The authors report that the 25th percentile commuter loses 98 cents per trip.
3. The authors report that the 25th percentile commuter gains 26 cents per trip.

that separate job-rich Orange County and housing-rich Riverside County. As a result, the HOT lanes are ten miles long, without any intermediate entrances and exits, and the time savings for users is as much as ten to fifteen minutes a trip. Such an enormous time savings encourages more than 40 percent of commuters on SR-91 to carpool, a much higher percentage than found on roads with shorter HOV or HOT lanes. If the authors had used a more typical commuter highway as their example, their scheme of different toll rates on general- and special-purpose lanes might have outperformed the HOT lanes.[4]

The authors also may have overestimated the benefits of HOT lanes to the extent that they overestimated the preferences that commuters have for the intrinsic characteristics of certain options rather than the savings in cost, travel time, and travel reliability that those options offer. The authors acknowledge that consumer utility increases slightly in their model as the number of options increases, even if those options do not actually offer more variety. But it also appears that the intrinsic preferences for certain options, such as owning a transponder or traveling in the exclusive lanes, are estimated to be high as well. This may give HOT lanes an important advantage over congestion pricing in the model simply because HOT lanes offer commuters a wide variety of options. Unfortunately, the econometric techniques the authors must use to combine different surveys with different types of data are so complicated that it is difficult to tell whether the intrinsic preferences for options have been overstated.[5]

In sum, Small, Winston, and Yan's basic point is probably more important than their specific example of congestion pricing on SR-91 makes it seem. On a more typical highway, with shorter lanes and commutes, differential pricing on the special- and general-purpose lanes might have beaten HOT lanes. But the success of the HOT lanes on SR-91 supports, rather than refutes, their argument that offering more options improves public policy where preferences are heterogeneous. HOT lanes offer more variety than congestion pricing, so one should not be surprised that they do reasonably well.

4. The authors simulate a road with lower HOV shares but they do so by adjusting down the intrinsic preference for HOV rather than by changing the characteristics of the road that give it such a high HOV share.

5. In an earlier version of their paper, the authors reported how much of the estimated consumer surplus per commuter was due to increases in the number of alternatives, transponder and special lane preferences, and differences in toll, time, or reliability. Roughly half of the welfare gain for many options was due to increases in the number of alternatives and transponder and special lane preferences, while the remainder was due to differences in toll, time, or reliability.

References

Brownstone, David, and Kenneth Train. 1999. "Forecasting New Product Penetration with Flexible Substitution Patterns." *Journal of Econometrics* 89 (1–2): 109–29.

Calfee, John, and Clifford Winston. 1998. "The Value of Automobile Travel Time: Implications for Congestion Policy." *Journal of Public Economics* 69 (1): 83–102.

Calfee, John, Clifford Winston, and Randolph Stempski. 2001. "Econometric Issues in Estimating Consumer Preferences from Stated Preference Data: A Case Study of the Value of Automobile Travel Time." *Review of Economics and Statistics* 83 (4): 699–707.

Choi, Ki-Hong, and Choon-Geol Moon. 1997. "Generalized Extreme Value Model and Additively Separable Generator Function." *Journal of Econometrics* 76 (1–2): 129–40.

Dahlgren, Joy. 1998. "High Occupancy Vehicle Lanes: Not Always More Effective than General Purpose Lanes." *Transportation Research A* 32 (2): 99–114.

De Palma, André, and Robin Lindsey. 2004. "Congestion Pricing with Heterogeneous Travelers: A General-Equilibrium Welfare Analysis." *Networks and Spatial Economics* 4 (2): 135–60.

Hensher, David A. 2001. "The Valuation of Commuter Travel Time Savings for Car Drivers in New Zealand: Evaluating Alternative Model Specifications." *Transportation* 28 (2): 101–18.

Hess, Stephane, Michel Bierlaire, and John W. Polak. 2005. "Estimation of Value of Travel-Time Savings Using Mixed Logit Models." *Transportation Research A* 39 (2–3): 221–36.

Jiang, Meilan, and Takayuki Morikawa. 2004. "Theoretical Analysis on the Variation of Value of Travel Time Savings." *Transportation Research A* 38 (8): 551–71.

Lee, Lung-fei. 1992. "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models." *Econometric Theory* 8: 518–52.

———. 1995. "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models." *Econometric Theory* 11: 437–83.

Manski, Charles F., and Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica* 45 (8): 1977–88.

McFadden, Daniel, and Kenneth Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15 (5): 447–70.

Mohring, Herbert. 1999. "Congestion." In *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer,* edited by J. Gómez-Ibáñez W. Tye, and C. Winston, pp. 181–222. Brookings.

Orski, C. Kenneth. 2001. "Carpool Lanes—An Idea Whose Time Has Come and Gone." *TR News* 214 (May–June): 24–28.

Poole, Robert W., Jr., and Ted Balaker. 2005. "Virtual Exclusive Busways: Improving Urban Transit While Relieving Congestion." Policy Study 337. Los Angeles: Reason Foundation.

Pratt, Richard H., and others. 2000. *Traveler Response to Transportation System Changes: Interim Handbook.* Transportation Cooperative Research Program (www4.nationalacademies.org/trb/crp.nsf/All+Projects/TCRP+B-12 [May 27, 2005]).

Santos, Georgina, ed. 2004. *Road Pricing: Theory and Evidence.* Oxford, United Kingdom: Elsevier Press.

Schrank, David, and Tim Lomax. 2005. *The 2005 Urban Mobility Report.* Texas Transportation Institute, Texas A&M University System, College Station, Texas.

Small, Kenneth A. 1992. *Urban Transportation Economics 51, Fundamentals of Pure and Applied Economics Series.* Chur, Switzerland: Harwood Academic Publishers.

Small, Kenneth A., and Clifford Winston. 1999. "The Demand for Transportation: Models and Applications." In *Essays in Transportation Policy and Economics: A Handbook in Honor of John R. Meyer,* edited by Gómez-Ibáñez, Tye, and Winston, pp. 11–55.

Small, Kenneth A., Clifford Winston, and Carol Evans. 1989. *Road Work: A New Highway Pricing and Investment Policy.* Brookings.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005a. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73 (4): 1367–82.

———. 2005b. Supplement to "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica Supplementary Material* 73 (4) (www.econometricsociety.org/suppmatlist.asp).

Small, Kenneth A., and Jia Yan. 2001. "The Value of 'Value Pricing' of Roads: Second-Best Pricing and Product Differentiation." *Journal of Urban Economics* 49 (2): 310–36.

Steimetz, Seiji S. C., and David Brownstone. 2005. "Estimating Commuters' 'Value of Time' with Noisy Data: A Multiple Imputation Approach." *Transportation Research B* 39 (10): 865–89.

Sullivan, Edward, and others. 2000. *Continuation Study to Evaluate the Impacts of the SR 91 Value-Priced Express Lanes: Final Report.* Department of Civil and Environmental Engineering, California Polytechnic State University at San Luis Obispo (December) (ceenve.ceng.calpoly.edu/sullivan/SR91).

Train, Kenneth E. 2003. *Discrete Choice Methods with Simulation.* Cambridge University Press.

U.S. Bureau of Public Roads. 1964. *Traffic Assignment Manual.* Washington.

Verhoef, Erik T., Peter Nijkamp, and Piet Rietveld. 1996. "Second-Best Congestion Pricing: The Case of an Untolled Alternative." *Journal of Urban Economics* 40 (3): 279–302.

Verhoef, Erik T., and Kenneth A. Small. 2004. "Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users." *Journal of Transport Economics and Policy* 38 (1): 127–56.

Yan, Jia, Kenneth Small, and Edward Sullivan. 2002. "Choice Models of Route, Occupancy, and Time-of-Day with Value Priced Tolls." *Transportation Research Record* 1812: 69–77.