PERGAMON

Transportation Research Part A 37 (2003) 499–517

TRANSPORTATION
RESEARCH
PART A

www.elsevier.com/locate/tra

Service quality—developing a service quality index in the provision of commercial bus contracts

David A. Hensher *, Peter Stopher, Philip Bullock

Faculty of Economics and Business, Institute of Transport Studies, University of Sydney, Sydney, NSW 2006, Australia

Received 26 September 2001; received in revised form 6 February 2002; accepted 21 November 2002

Abstract

The measurement of service quality continues to be a challenging research theme and one of great practical importance to service providers and regulatory agencies. The key challenges begin with the identification of the set of potentially important dimensions of service quality perceived by passengers, current and potential. We then have to establish a way of measuring each attribute and identifying their relative importance in the overall calculation of satisfaction associated with existing service levels. Once a set of relevant attributes has been identified, this information can be integrated into programs such as monitoring and benchmarking, and even in contract specification. This paper, building on earlier research by the authors, investigates ways of quantifying service quality and comparing the levels within and between bus operators. The importance of establishing suitable market segments and the need to scale the service quality index for each operator to make meaningful comparisons is highlighted.

Keywords: Service quality; Bus reform; Stated preference; Choice modelling

1. Background

There is an extensive literature (Fielding et al., 1985) on measuring the cost efficiency and cost effectiveness of bus services and operations. A major data input is the level of service output, typically measured on the demand side by annual passenger trips or passenger kilometres and on

* Corresponding author. Tel.: +61-2-9351-0071; fax: +61-2-9899-6674.
E-mail addresses: davidh@its.usyd.edu.au (D.A. Hensher), peters@its.usyd.edu.au (P. Stopher), philipb@its.usyd.edu.au (P. Bullock).

0965-8564/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0965-8564(02)00075-7

the supply side by vehicle kilometres. As aggregate indicators of total output, these measures implicitly assume homogeneity of service quality. Passengers, however, evaluate services in many ways that may not be systematically associated with the amount of use of the service; indeed it is unclear whether aggregate passenger kilometres can be a proxy for differences in passenger satisfaction across bus segments.

Several studies have since refocused on the measurement of service quality, investigating the role of trade-off methods such as stated preference (SP) (e.g. Prioni and Hensher, 2000; Hensher, 1991 and Swanson et al., 1997) and univariate procedures that rate individual service items on a satisfaction scale (Cunningham et al., 1997). Although a passenger may perceive specific aspects of service quality as either positive or negative, we assume that the overall level of passenger satisfaction is best measured by how an individual evaluates the total package of services offered. Appropriate weights attached to each service dimension will reveal the strength of positive and negative sources of overall satisfaction. The SP paradigm enables us to develop preference formulae for a large number of service level scenarios, which can be implemented at the bus business level to establish operator-specific indicators of service delivery quality and effectiveness. The resulting satisfaction (utility) indicators obtained from the SP experiments measure the expected utility that a passenger obtains from the current levels of service and how this might change under alternative service level regimes. [1]

In 1999, the Institute of Transport Studies (ITS) began researching ways the bus and coach industry in New South Wales (Australia) might capture customer satisfaction with service levels (Prioni and Hensher, 2000; Hensher and Prioni, 2002). The intention was to provide insights into how quality could be built into a possible future government performance assessment regime, including calculating value for money in commercial bus contracts. It would also provide insights into the effectiveness of service levels from a passenger viewpoint and identify which service aspects are working best and which need more improvement. ITS undertook a pilot program in which an on-board customer survey was undertaken with the support of 25 operators, focusing on a current trip and seeking information on passenger perceptions of service levels on 13 predetermined attributes. Stated choice (SC) methods were used, in which a sample of passengers were asked to choose their most preferred package from a number of alternative packages of service levels based on these attributes. Multinomial logit (MNL) models were estimated to establish the relative weights attached to the statistically significant attributes, representing the contribution of each service attribute to the calculation of an overall service quality index (SQI). The pilot program showed the value of SQI as a way to capture customer perceptions of service quality.

In 2000, we embarked on the development phase. Two key features were identified that needed more attention: selection of service segments within an operator's domain, and a carefully structured sampling plan. This paper presents the findings of this development phase. One major public operator and one major private operator were invited to participate and asked to propose service segments. A total of nine service segments were surveyed in this current round, sufficient to

[1] Given the heterogeneity of the population of bus passengers, segment-specific service quality indicators can be identified.

establish a benchmarking capability for ongoing monitoring for each segment and, through aggregation, for each operator.

We begin with an overview of the data requirements for quantifying SQI, including the selection of the attributes and the role of SC methods. The sampling plan is then presented. The logistics of data collection are described followed by a summary of the sample responses and a profile of the data on passenger perceptions of current service levels. Next, we describe the statistical models that establish the weights associated with each attribute in each service segment for each operator. Because we wish to benchmark each operator's market segment against the other segments, we introduce some specific details of how the statistical analysis is undertaken. In brief, because the relative importance of an attribute in a segment is scaled for comparability within the segment, to be able to undertake comparisons between segments we have to rescale the weights. The SQI measures are then calculated for each market segment with a comparison between each segment in terms of the overall SQI and its constituent attributes. The paper concludes with a summary of major findings.

2. Data requirements and attribute selection for service quality measurement

2.1. The stated preference approach

The task is to develop an SQI that can be incorporated into a performance assessment regime that measures service effectiveness meaningfully from a passenger perspective. Such an index should be able to be decomposed into its constituent sources of passenger satisfaction. It should also map into an aggregate demand-side indicator of passenger output to establish the role of the latter as a practical approximation of the social welfare significance of bus service levels.

With a complex disaggregation of service quality, data reflecting the experience from an existing trip alone, referred to as revealed preference (RP) data, are usually inappropriate. There is potentially too much confounding in RP data. Furthermore some attributes of interest (e.g., air conditioning, low floor entry) may not exist today on many urban buses, so their influence cannot be determined.

SP methods provide the data richness required for quantifying an SQI, involving an SC experiment in which we systematically vary combinations of levels of each attribute to reveal new opportunities relative to existing service levels (Hensher, 1994; Hensher et al., 1999; Louviere et al., 2000). The attributes must be anchored to current experience, so that respondents can understand and relate to the attribute levels in a realistic way (Stopher, 1998). It is then important to create the other possible levels as reasonable variations on either side of current experience. Failure to do this leads to respondents providing poor quality and inappropriate responses, as they try to relate to attribute levels that are totally outside their experience and sometimes difficult to imagine (Louviere et al., 2000). Through the experimental design approach, we survey a sample of travellers making choices between the current and other trip attribute level bundles. This approach is capable of separating out the independent contributions of each service component and hence is capable of providing an SQI that is a rich representation of the statistically significant sources of service (dis)utility.

Table 1
Attributes and attribute levels in the sp experiment

	Attribute	Level 1	Level 2	Level 3
1	Bus travel time	25% less	Same	25% more
2	Bus fare	20% less	Same	20% more
3	Ticket type	Cash fare	Pre-purchased bus-only 10-trip ticket or weekly	Integrated (bus and other mode)
4	Buses per hour at this bus stop (i.e., frequency)	50% more service	Same as now	50% less service
5	Time of arrival at bus stop	On time	5 min late	10 min late
6	Time walking to bus stop	Same	An extra 5 mins	An extra 10 mins
7	Seat availability on bus	Seated all the way	Stand part of the way	Stand all of the way
8	Information at bus stop	Timetable and map	Timetable, no map	No timetable, no map
9	Access to bus	Wide entry, no steps	Wide entry, 2 steps	Narrow entry, 4 steps
10	Bus stop facilities	Seats only	Seats under cover	No seat or shelter
11	Temperature on bus	Too hot	Just right	Too cold
12	Driver attitude	Very friendly	Friendly enough	Generally unfriendly
13	General Cleanliness on board	Very clean	Clean enough	Not clean enough

2.2. Defining the empirical setting and the SP experiment

To help select attributes for SQI, we undertook an extensive literature review and a survey of bus operators with a wealth of experience on what customers look for in a good service (Prioni and Hensher, 2000). We also benefited from the earlier pilot study (Hensher and Prioni, 2002). Together with extensive discussions during the development stage with key bus operators in Sydney, we concluded that 13 attributes describe the major dimensions of service quality from a user's perspective. [2] The range of levels selected for each attribute are shown in Table 1. The attribute ranges were selected in consultation with the operators as representative of achievable variability. The attributes treated as continuous for the SP alternatives were relative to the current trip levels (e.g. travel time). The classificatory attributes (e.g. temperature on bus) included the level selected as the reference level for the current trip (from the levels offered in Table 1). The current level refers to the level associated with the trip in progress when the onboard survey was implemented. These data were obtained from passengers prior to completing the SP experiment).

Through a formal statistical design, the attribute levels are combined into bus packages before being translated into a survey form. The full factorial design consists of 3^{13} combinations of the three levels of the 13 attributes. To produce a practicable and understandable design for respondents, we restricted the number of combinations to 81 choice sets using a fractional design that permits reduction of the number of bus packages, without losing important statistical information (Louviere et al., 2000). A pretest showed that respondents were able to evaluate consistently three choice sets resulting in 27 different survey forms. To allow for a rich variation in the combinations of attribute levels to be evaluated as service packages in the SP experiment, each bus

[2] These 13 attributes are not the same set as those evaluated in the pilot.

operator received eight sets of 27 different survey forms (i.e., 216 forms) and instructions on how to organise the survey.

2.3. Sampling strategy

The overall sampling plan was to distribute approximately 500 surveys on each of three segments (route types) from each of three depots, totalling 4500 surveys. In addition, each of peak and off-peak runs were to be surveyed in each segment. The sample design was a multistage sample where the first stage was a stratified sample of routes within segments, the second stage a stratified sample of bus runs within sampled routes for each of peak and off-peak, and the third stage a census of riders on selected runs. The third stage census makes it easier to administer in the field, because the surveyors do not have to perform any type of selection process, and cannot introduce a bias into the procedure. It also reduces the addition of further sampling error at this stage. In this paper, the geographical service segments are assigned the identifiers S1 to S9. Some segments were CBD-based services while others were local and cross regional services serving rail stations and local centres.

Peak travel was sampled more heavily than off-peak, with 2700 surveys in the peak (7–9 am) and 1800 in the off-peak (10 am–2 pm). Within each depot, the surveys were assigned equally to each segment. Thus, each segment was to have 500 surveys distributed, with 300 in the peak and 200 in the off-peak, rounded upwards to allow complete runs to be sampled. Each route in these segments was sampled approximately equally, as far as average ridership per run allowed. For three segments, there were only two or three routes in each segment. In the six other segments, there were too many routes to allow even one run per route to be included in the sample, so a simple random sample of routes and runs was chosen from each segment, until the desired expected ridership was reached.

2.4. Logistical issues in data collection

The coordinator for each operator was briefed the week before the survey began and provided with the survey forms (sorted by bus segment) and the sampling rules. We sorted and allocated the 27 sets of survey forms, to ensure an equal distribution within each segment. The survey was undertaken in the last week of November and first week of December in 2000. One operator used their own senior staff to distribute and collect the forms and the other operator hired a survey firm for this task.

Although specific bus runs were provided by the sampling plan, the most important compliance condition was that the appropriate number of forms was distributed within each route within each segment for each of the two time periods (7–9 am and 10 am–2 pm). For certain segments, operators were concerned about crowding conditions hampering the distribution of the forms. Shortage of interviewers on these segments combined with a higher number of bus runs meant that the required number of surveys could not always be circulated. In addition, there were often few customers on off-peak services as well as more elderly passengers for which there was a high rate of refusal. For other segments crowding was not a concern and so the full number of survey forms could be distributed using a similar number of interviewers as on the more crowded segments. For the off-peak components of service segments S2 and S3 the actual number of forms distributed exceeded the planned distribution, though this did not offset the shortfall in peak distributions. Overall, sufficient forms were returned to undertake the segment-specific analysis and determination of SQI.

3. Sample response

Table 2 presents response rates in several ways for each operator, segment and time of day. First is the planned distribution, followed by the number returned. The survey instrument comprised two double-sided A4 pages. [3] The first side had 25 questions about the current trip and the respondent. The attribute data sought was identical to that offered in the SP packages except for the service levels offered. Each of the remaining three sides set out one choice set for the SP experiment. The returns are grouped into four categories: front page details (RP) incomplete, RP completed only, RP plus one experiment only completed, RP plus two experiments completed, and entire form completed. Surveys that have the RP and at least one SP experiment completed are useful in statistical analysis: usable surveys are the sum of the latter three categories.

The actual number handed out is not known. This was partly due to the method used to administer the survey. When respondents returned blank forms, interviewers handed the same form to the next passenger. Logistical issues such as these will need to be re-assessed for future surveys. The number handed out was also not always the same as designed, because there were runs on which there were fewer passengers than expected.

These response rates are considered good for an on-board bus survey, where response rates are often as low as 15–25%. Only three segments showed a response rate against the planned distribution below 25%. In addition, completion rates of the surveys are considered good for on-board surveys indicating that the instrument is working well and that response to the survey has been positive. However, although interviewers were instructed not to give surveys to schoolchildren, they did so, which inflated the numbers somewhat, while providing surveys that must be excluded.

3.1. Data profile

Table 2 provides a profile of the socioeconomic composition of sampled passengers. For each categorical person characteristic we provide the distribution of category membership; for each continuous variable we provide the mean, standard deviation and range. In Table 3, totals are the number of respondents, excluding school children, who completed one or more SP experiments.

[3] Although laptop-based SC experiments are generally preferred and indeed it is the standard method used by the authors in most studies, it was not possible to undertake the current on-board survey using laptops. Interviewers would have been required and the cost would have been well beyond the available budget of the bus operators. We have recently developed an internet.based survey instrument which will enable operators to undertake service quality surveys at little expense prior to analysis. The downside is the preservation of a representative sample.

Table 2
Response rates by segment

Segment	Time of day	Planned dist	Status of survey form completion					Total responses	Total usable responses (% of total responses)	Usable responses as % of planned dist
			RP incom-plete	RP only	RP+ SP1	RP+ SP1+ SP2	RP+ SP1+ SP2+ SP3			
S1	Peak	343	21	31	10	15	86	163	111 (68)	32
	Off-peak	200	1	50	12	9	84	156	105 (67)	53
S2	Peak	241	2	27	4	11	60	104	75 (72)	31
	Off-peak	158	26	51	9	12	90	188	111 (59)	70
S3	Peak	328	69	23	5	11	59	167	75 (45)	23
	Off-peak	216	59	69	17	13	74	232	104 (45)	48
S4	Peak	310	10	39	12	13	68	142	93 (65)	30
	Off-peak	203	9	47	8	6	64	134	78 (58)	38
S5	Peak	322	13	45	16	19	98	191	133 (70)	41
	Off-peak	210	8	72	12	14	86	192	112 (58)	53
S6	Peak	302	13	51	15	11	49	139	75 (54)	25
	Off-peak	193	8	52	16	16	35	127	67 (53)	35
S7	Peak	337	0	20	6	16	95	137	117 (85)	35
	Off-peak	224	1	45	9	8	38	101	55 (54)	25
S8	Peak	303	2	6	3	5	43	59	51 (86)	17
	Off-peak	220	4	42	15	8	46	115	69 (60)	31
S9	Peak	297	2	12	5	4	28	51	37 (73)	12
	Off-peak	214	2	17	1	9		29	10 (34)	5

Age and income were transformed from categorical data to continuous data (for modeling purposes). Lowest (18 and under) and highest age categories (65 and over) were recoded into 18 and 70 respectively. The lowest (under $12,000) and highest (over $80,000) income categories were recoded into $12,000 and $100,000 respectively.

The attributes associated with the SP design are summarised in Tables 4 and 5 but only for the levels associated with the current trip (RP levels). These are useful, because they are the basis of the input data used to calculate the SQI. A very high proportion of the sample had a seat all the way, suggesting either that there were few standing passengers or that standing passengers found it difficult to complete the surveys. This will need to be checked in future surveys, because it may be a potential source of bias. The other attributes have a good spread of responses in at least two of the three levels. On-board temperature is "just right" for about 71% of passengers; the balance see it as too hot (and rarely too cold). Buses tend to be either very clean or clean enough and drivers tend to be very friendly or friendly enough.

On-time running (i.e., unreliability) shows buses arriving up to 25 min early or late with an average in the 0.2–3 min range. The majority of buses however arrived between 3 min early and

Table 3
Socioeconomic data by segment

Variable (%)	Segment									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	All
Gender										
Female	56.0	62.4	60.3	66.1	57.1	54.2	52.9	65.0	46.8	58.6
Male	40.7	34.4	36.3	33.3	40.8	36.6	45.9	33.3	51.1	38.5
Missing	3.2	3.2	3.4	0.6	2.0	9.2	1.2	1.7	2.1	2.9
Total (N)	216	186	179	171	245	142	172	120	47	1478
Main occupation (%)										
Employed full time	38.4	40.9	41.3	56.7	43.3	37.3	68.6	46.7	51.1	46.5
Student	18.5	23.7	17.9	24.0	27.8	14.1	11.6	18.3	23.4	20.2
Looking for work	11.6	5.9	7.3	4.1	2.4	2.1	1.7	3.3	4.3	5.0
Retired or pensioner	11.1	5.9	13.4	3.5	13.9	38.7	10.5	19.2	6.4	13.4
Home duties	10.6	12.4	13.4	3.5	3.3	2.8	1.7	7.5	4.3	6.9
Other	9.3	9.7	5.6	6.4	9.0	4.2	4.7	5.0	8.5	7.1
Missing	0.5	1.6	1.1	1.8	0.4	0.7	1.2		2.1	0.9
Total (N)	216	186	179	171	245	142	172	120	47	1479
Age (years)										
Mean	33.5	31.7	36.2	34.4	35.7	43.7	37.7	39.6	35.3	36.1
Standard deviation	14.6	13.5	16.0	13.4	16.1	19.6	15.3	16.8	14.7	15.9
Missing N	1	2		2	3	1	2		1	12
N	216	186	179	171	245	142	172	120	47	1497
Income (dollars)										
Mean	28,590.4	31,664.9	28,935.8	37,918.5	30,573.6	28,652.2	48,436.0	32,406.9	36,584.8	33,294.8
Standard deviation	16,688.5	22,147.3	17,014.7	20,518.9	19,980.2	18,149.8	24,316.4	19,190.7	22,198.3	20,845.2
Missing		2		2	3	2	2		1	13
N	216	186	179	171	245	142	172	120	47	1479

3 min late. This is one attribute on which the regulator places a great deal of importance (because it is relatively easy to measure external from the passenger); as shown later, it is also a statistically significant influence for the passenger.

Each passenger was given a choice set of three alternatives to evaluate (the current and two SP designed packages selected from the 81 available sets. They evaluated them and chose one. This was repeated a total of three times. Packages A and B are unlabelled (or generic) alternatives defined by the bundle of attribute levels and as such each package has no branding. Overall, however, 50.6% of passengers from one operator chose their current bus package (choice C), with 46% of passengers from the other operator choosing their current bus package.

Table 4
RP data by segment (categorical variables)

Attribute (%)	Segment									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	All
Seat availability on bus										
Seat all the way	95.8	97.3	96.6	92.4	91.4	99.3	94.8	95.0	100.0	95.3
Stand for part of the way	3.2	2.7	2.8	5.8	8.6	0.7	3.5	5.0		4.1
Stand all the way	0.9		0.6	1.8			1.7			0.6
Total (N)	216	186	179	171	245	142	172	120	47	1478
Bus stop facilities										
Seats only	6.9	10.2	15.6	36.8	31.4	45.8	32.6	24.2	8.5	24.1
Seats under cover	52.8	30.1	49.7	48.5	54.3	37.3	58.1	62.5	74.5	49.9
No seats or shelter	40.3	59.7	34.6	14.6	14.3	16.9	9.3	13.3	17.0	26.0
Total (N)	216	186	179	171	245	142	172	120	47	1478
Information at bus stop										
Timetable and map	9.7	4.8	5.0	25.1	27.3	18.3	18.6	6.7	8.5	14.8
Timetable but no map	32.9	30.6	30.7	49.7	48.6	40.1	45.3	18.3	27.7	37.7
No timetable and no map	57.4	64.5	64.2	25.1	24.1	41.5	36.0	75.0	63.8	47.5
Total (N)	216	186	179	171	245	142	172	120	47	1478
Access to the bus										
Wide entry, no steps	8.8	8.1	24.6	19.9	37.6	4.2	45.9	49.2	29.8	24.5
Wide entry, two steps	74.5	78.0	57.0	70.2	55.1	77.5	44.2	41.7	61.7	62.8
Narrow entry, four steps	13.4	7.5	12.8	4.7	4.1	14.1	5.8	5.8	6.4	8.4
Other	3.2	6.5	5.6	5.3	3.3	4.2	4.1	3.3	2.1	4.3
Total (N)	216	186	179	171	245	142	172	120	47	1478
Temperature on bus										
Just right	60.6	68.8	72.6	77.8	87.3	45.1	64.0	76.7	89.4	70.6
Too hot	38.9	31.2	26.3	19.3	6.1	52.8	32.6	22.5	10.6	27.1
Too cold	0.5		1.1	2.9	6.5	2.1	3.5	0.8		2.3
Total (N)	216	186	179	171	245	142	172	120	47	1478
Cleanliness of bus										
Very clean	24.1	32.3	27.9	24.0	29.4	26.1	33.7	34.2	36.2	29.0
Clean enough	70.4	65.6	69.8	70.8	65.3	69.7	57.0	59.2	53.2	65.8
Not clean enough	5.6	2.2	2.2	5.3	5.3	4.2	9.3	6.7	10.6	5.2
Total (N)	216	186	179	171	245	142	172	120	47	1478
Friendliness of driver										
Very friendly	36.6	42.5	40.8	15.2	30.6	34.5	18.0	24.2	34.0	30.9
Friendly enough	59.7	54.3	55.3	78.9	62.4	61.3	77.9	72.5	66.0	64.7
Very unfriendly	3.7	3.2	3.9	5.8	6.9	4.2	4.1	3.3		4.4
Total (N)	216	186	179	171	245	142	172	120	47	1478

Table 5
RP data by segment (continuous variables)

Attribute (%)	Segment									
	S1	S2	S3	S4	S5	S6	S7	S8	S9	All
Time to get to bus stop (min)										
Minimum	0.2	0.5	0.5	0.2	1.0	0.5	1.0	0.5	1.0	0.2
Maximum	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	25.0	30.0
Mean	5.5	5.0	5.2	5.4	8.3	7.2	7.4	7.5	6.0	6.4
Standard Deviation	5.4	4.9	5.6	5.1	6.2	5.9	5.4	6.2	5.8	5.7
N	216	186	179	171	245	142	172	120	47	1478
On time unreliability of bus (min)[a]										
Minimum	−12.0	−0.18	−0.24	−20.0	−25.0	−23.0	−20.0	−16.0	−23.0	−25.0
Maximum	17.0	10	19	20.0	19.0	20.0	20.0	20.0	15.0	20.0
Mean	1.9	0.76	1.2	3.51	1.1	1.2	0.8	0.2	0.4	1.4
Standard deviation	4.3	4.3	4.7	5.34	4.9	4.9	4.8	4.3	6.8	4.9
N	216	186	179	171	245	142	172	120	47	1478
Number of buses in 1 h interval										
Minimum	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Maximum	12.0	12.0	5.0	12.0	12.0	12.0	12.0	10.0	9.0	12.0
Mean	3.6	2.7	2.0	4.9	4.8	3.6	5.2	2.3	2.8	3.7
Standard deviation	1.7	1.4	0.7	2.4	2.7	2.0	2.5	1.2	1.6	2.3
N	216	186	179	171	245	142	172	120	47	1478
Travel time on bus (min)										
Minimum	5.0	5.0	5.0	6.0	5.0	5.0	7.0	5.0	4.0	4.0
Maximum	60.0	45.0	30.0	60.0	60.0	60.0	60.0	60.0	30.0	60.0
Mean	17.3	17.9	13.8	30.3	24.8	19.7	32.2	22.7	17.1	22.1
Standard deviation	9.1	9.0	4.8	13.1	13.9	10.8	10.1	11.4	7.5	12.2
N	216	186	179	171	245	142	172	120	47	1478
Cost of current one way fare (dollars)										
Minimum	0.10	0.07	0.32	0.52	0.52	0.52	0.60	0.52	0.39	0.07
Maximum	5.85	6.34	6.50	5.00	7.20	7.20	9.00	9.00	3.75	9.00
Mean	1.97	1.95	1.77	1.96	2.03	1.66	2.26	1.90	1.69	1.94
Standard deviation	0.84	0.89	0.98	0.91	1.36	1.17	1.24	1.11	0.79	1.08
N	216	186	179	171	245	142	172	120	47	1478

[a] Number of minutes that the bus was late. Negative values refer to the bus running early.

4. Statistical analysis to quantify service quality

4.1. SQI and importance weights

The derivation of SQI requires statistical estimation of models that reveal the importance weights attached to each attribute by the sample of passengers in each segment. The perceptions of passengers relative to the levels of each attribute as experienced in a current trip, and the levels offered in each SP package, together with the choice of the preferred trip package provide the necessary information to identify the importance weights. When the weights are identified, we have to multiply each attribute level associated with the current trip by the relevant weight and sum these calculations across all attributes to produce the SQI for each sampled passenger. An average across all sampled passengers using a specific segment provides the segment SQI, which measures overall perceived satisfaction with existing service levels. Each segment will have an overall SQI as well as information on the contribution of each attribute to that SQI. The latter is very useful in helping the operator gain an understanding of what are the main positive and negative influences on the overall level of passenger satisfaction with current services.

4.2. Benchmarking and discrete choice modelling

The MNL model identifies the importance weights (Louviere et al., 2000). This simple method to obtain the importance weights has one limitation when the interest is in benchmarking SQI across geographical service segments. The desire to have separate models for each segment is linked to establishing unique importance weights for each attribute within each segment. We could naively pool the data across all segments and treat the importance weights for each attribute as the same (Prioni and Hensher, 2000). In principle this is quite acceptable if there are no statistically significant differences in the levels of the importance weights across the segment samples of passengers. However, it must be demonstrated rather than assumed.

For benchmarking we need to ensure that the SQI measures are comparable between segments. A discrete choice model is structured so that the information on the importance of each attribute is relative within a model. For example, if the importance weight for unreliability in segment [4] 1 is −0.4 and for bus fare it is −0.04, then we can compare these two weights and conclude that the unreliability weight per unit of unreliability is 10 times more valuable than the bus fare weight per unit of fare. If, in segment 2, the unreliability weight from a separately estimated model is −0.2, we cannot conclude that unreliability per unit is valued at twice the rate in segment 1 as segment 2, because each separately estimated model has a different scale structure for comparing the importance weights.

Specifically, the MNL model derives importance weights with two components—scale and taste. Scale is derived from the underlying assumptions of the error structure. The MNL model assumes that this scale is the same across the alternatives being evaluated (i.e., current trip levels and the two SP packages) and can be set to 1.0. While this assumption can hold within a segment, we cannot assume that it holds across segments. If we assumed this, then we could pool the data for each segment and treat the scale as 1.0 for all alternatives associated with all segments. Because we have no way of knowing it is true before we test the assumption, we have to redefine the structure of the model to be estimated so it can reveal the extent of differences in scale (if they exist) when we pool the data. We must pool the data because we want to undertake benchmarking and must ensure that the importance weights (and hence segment SQIs) are directly comparable.

To account for potential scale differences, each segment is treated as having three alternatives (current plus two SP packages), which are different because of scale differences. We then have 27 alternatives. The structure is shown in Fig. 1, where each respondent provides information to one branch in the tree structure and each branch is a segment (S1,…, S9), revealing the scale

[4] Segments 1, 2 are general and hypothetical terms here; this discussion does not refer to specific results of this experiment.

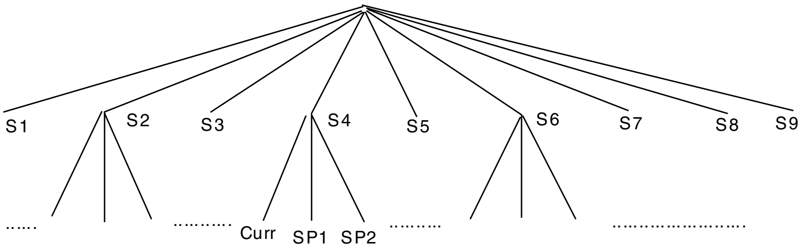

Fig. 1. Nesting structures used in model estimation to permit comparisons of SQI between nine segments.

differences empirically. Previous research (Hensher and Bradley, 1993) shows how the scale parameter can be identified using this procedure. Essentially we use the nested logit structure as a 'trick' to reveal differences in scale since we are pooling data across nine separate samples drawn from the nine geographical segments. We normalise the scale value for one segment (by setting it to 1.0—segment four) and allow it to be free for the other eight segments.

4.3. SQI model

The final nested model system is summarised in Table 6. The dependent variable is binary, where the chosen package is given the value of 1.0 and the other two non-chosen packages are given the value of zero. The model then establishes the statistically most efficient weights to explain the choices made. The focus of the survey is on the attributes themselves, without reference to any particular label describing any package. Each package is a combination of attributes and associated levels and is referred to as an unlabelled alternative. Hence all the weights attached to a specific attribute (e.g., travel time) are the same across the three packages. These weights can vary between segments, but where they are not statistically significantly different, they are constrained to be the same in the final model (Table 6). Eight scale parameters have been identified relative to the scale for segment 4 which is set equal to 1.0. The overall explanatory power of this highly non-linear model is very high (a pseudo-R^2 of 0.69). [5] For a linear model this is close to 0.9.

Some attributes are not statistically significant in all segments and hence their contribution to SQI is not significant. The presence of strong support for a specific statement (e.g., "this bus is very clean") does not indicate that this is an important issue in the passenger's overall satisfaction with the service package being offered. This choice approach, in which one evaluates the current trip levels and two alternative service packages, enables one to assess the role of each attribute in influencing the choice between service packages. Six variables are significant for all segments, namely one-way bus fare, seat all the way, stand part way, wide entry two steps, seat only at stop, and seat under cover. In addition, travel time is significant on all but S3, bus frequency on all but S9, access time to bus stop on all but S4 and S7, narrow 4 steps on two of the three segments where it was presented, and very clean bus on three of the four segments where it was presented. Unreliability

[5] We investigated a number of attribute interactions but they did not add significantly to the overall goodness-of-fit. However we wanted to keep the formula linear in order to simplify the process of excluding specific attributes where the regulatory focus on service quality might be limited to a few attributes (see Hensher and Prioni, 2002).

Table 6
The final model used to identify the importance weights and scale differences between segments for scheduled route services (school children on passes have been excluded)

Attribute	Segment importance and scale weights (*t*-value in brackets)[a]								
	S1	S2	S3	S4	S5	S6	S7	S8	S9
Travel time (mins)	−0.0333	−0.0346	−0.0249	−0.0440	−0.0396	−0.0356	−0.0280	−0.0272	−0.0362
	(−3.8)	(−3.2)	(−1.5)	(−4.9)	(−3.9)	(−3.2)	(−3.3)	(−2.7)	(−2.1)
One-way bus fare ($)	−0.6519	−0.7136	−0.7508	−0.5592	−0.6394	−0.5948	−0.6256	−0.5543	−0.5543
	(−4.5)	(−4.4)	(−4.0)	(−4.3)	(−4.6)	(−4.4)	(−4.2)	(−2.9)	(−2.9)
Unreliability (mins)	−0.0317	−0.0322	−0.0626	−0.0399	−0.0649	−0.0119	−0.0116	−0.1127	−0.1029
	(−1.8)	(−1.4)	(−1.7)	(−2.6)	(−3.3)	(0.5)	(−0.8)	(−3.9)	(−1.9)
Access time to bus stop (mins)	−0.0248	−0.0725	−0.0859	−0.0081	−0.0449	−0.0696	−0.0128	−0.0567	−0.0768
	(−2.0)	(−3.9)	(−3.4)	(−0.8)	(−3.4)	(−3.4)	(−1.1)	(−3.6)	(−2.7)
Bus frequency (/hr)	0.0923	0.0840	0.2729	0.0490	0.0858	0.1187	0.0869	0.1440	0.0523
	(3.0)	(2.0)	(2.8)	(2.0)	(2.6)	(2.2)	(2.8)	(2.9)	(0.6)
Seat all way (1,0)	0.6529	0.6661	0.5159	0.4380	0.4622	0.5310	0.7734	0.3560	0.9531
	(3.8)	(3.0)	(2.5)	(3.1)	(2.8)	(2.1)	(4.7)	(1.9)	(2.0)
Stand part way (1,0)				0.2367	0.2367	0.2367	0.2367	0.2367	0.2367
				(2.5)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)
No timetable, no map (1,0)	−0.1850	−0.4216		−0.1372		−0.2464	−0.2913	−0.2033	−0.1210
	(−1.4)	(−2.3)		(−1.1)		(−1.5)	(−1.9)	(−1.2)	(−0.5)
Narrow 4 steps (1,0)	−0.4455	−0.1535					−0.5709		
	(−2.7)	(−0.8)					(−3.1)		
Wide entry 2 steps (1,0)	−0.5124	−0.4899					−0.5748		
	(−3.2)	(−2.7)					(−3.3)		
Seat only at stop (1,0)	0.6102	0.6102	0.6102	0.1851	0.1851	0.1851	0.1851	0.1851	0.1851
	(4.2)	(4.2)	(4.2)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)
Seat under cover at the bus stop (1,0)	0.6102	0.6102	0.6102	0.1851	0.1851	0.1851	0.1851	0.1851	0.1851
	(4.2)	(4.2)	(4.2)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)	(2.5)
Very clean bus (1,0)		0.3228			0.3228	0.2262		0.3228	
		(2.9)			(2.9)	(1.7)		(2.9)	
Very friendly driver (1,0)		0.1704	0.1704	0.2089	0.2263			0.2263	
		(1.4)	(1.4)	(1.7)	(1.9)			(1.9)	
VTTS ($/h)	3.06	2.92	1.99	4.72	3.72	3.59	2.68	2.94	3.92
No. of observations[b]	580	511	472	454	646	336	463	304	122
Scale value	0.9835	0.5019	0.6326	1.0000	0.7270	0.4212	1.065	1.0727	0.8370
	(4.6)	(3.8)	(4.4)	(fixed)	(4.7)	(3.0)	(5.6)	(4.4)	(3.2)
Log-likelihood					−3848.9				
Pseudo-R^2					0.69				

[a] Missing attribute weights mean that the attribute was too insignificant to report for the segment where it was highly non-significant.

[b] The minimum number of observations per respondent was 1 and the maximum was 3 (i.e., 3 or less SP experiments completed).

varied, sometimes being very significant (S4, S5, S8) and in other cases not significant. Very friendly driver was never significant, and no map, no timetable was significant only once.

Some attributes tend to be dominated by support on two levels, such as driver friendliness, bus cleanliness and getting a seat on the bus. For example, 95.6% of passenger responses described the driver as very friendly or friendly enough; and 94.8% of the passengers described the bus as either very clean or clean enough. Consequently the best specification of these attributes was achieved by setting the best level relative to the other levels. Hence "very clean" and "very friendly" are the only attributes in the model for bus cleanliness and driver friendliness. The interpretation of the importance weights is straightforward. A positive weight indicates that a very clean bus adds to utility the equivalent of its importance weight compared to a bus that is not perceived as very clean (i.e., is predominantly clean enough with a few passenger perceiving it as not clean enough).

The scale parameters, varying from 1.065 to 0.4212, are all statistically significant although there are quite a few segments where the scale parameters are statistically similar. Thus, we cannot pool all the data for each segment and treat the scale as 1.0 for all alternatives associated with all segments, but we could pool subsets of service segments. The SQI utility index has to be multiplied by this scale parameter to enable benchmarking. [6] Where the SQI is positive, a scale value less than 1.0 reduces the SQI value; where the SQI is negative a scale parameter less than 1.0 increases the SQI value. Table 7 highlights the implications of ignoring the scaling on rank ordering of segments. There is substantial re-ordering except for S4.

The implied values of travel time savings (VTTS) are informative. They vary from $2 to $4.72 per person hour. Bus users generally have lower VTTS than car and train users, with the values herein varying between 14% and 34% of the gross average wage rate of the sample. However, there is one important caveat to note—unlike previous studies we have separated out the in-vehicle time from the unreliability of travel time, which tends to otherwise inflate the mean VTTS for in-vehicle time.

4.4. Benchmarking service quality

The parameter estimates combined with the perceptions of service levels on each attribute associated with the current trip provide all the data necessary to derive the SQI measure for each segment. The products of each parameter and the associated attribute level across the sampled passengers are summarised in Table 7. We calculated the actual utility contribution of each attribute for each passenger, summed them in each segment and took the average. [7] The overall SQIs are shown in Fig. 2 and the contributions of each attribute are shown in Fig. 3.

[6] The values in Table 7 are calculated by multiplying the RP attribute levels by the appropriate weight in Table 6, summing across the sample of passengers in the segment and taking an average, and then multiplying by the scale parameter.

[7] A referee suggested that time and cost should not be included in the calculation of SQI, and if they should, then they should be normalized by distance. Although this is a very interesting insight and one that we have thought about a lot, we argue that all attributes are true contributors to service quality as promoted in this paper and the only way to properly place each attribute representing information on individual's preferences for alternative service levels is to include all statistically significant attributes. We would also argue that the utility or satisfaction associated with many of the attributes (notably all on-board attributes) varies with distance travelled and to exclude only time and cost on this reasoning is not valid. The focus is on passengers choosing a package of attribute levels as the basis of choosing one service over another. It is not focused on a preference for an individual attribute per se but on how the attributes are mixed (through packaging) in delivering an overall service level that is the basis of choosing a service. However a nice feature of the approach is that it is very easy to remove an attribute to re-benchmark. Results excluding time, cost and both attributes are available on request.

Table 7
SQI and its contributing components by segment (all scaled)

Variable	Segments								
	S1	S2	S3	S4	S5	S6	S7	S8	S9
Travel time	−0.573	−0.315	−0.218	−1.343	−0.719	−0.300	−0.965	−0.671	−0.534
One way bus fare	−1.273	−0.709	−0.859	−1.103	−0.953	−0.394	−1.412	−1.294	−0.773
Unreliability	−0.077	−0.030	−0.079	−0.159	−0.092	−0.010	−0.023	−0.150	−0.215
Bus frequency	0.327	0.113	0.350	0.236	0.303	0.182	0.481	0.360	0.125
Access time to bus stop	−0.135	−0.176	−0.286	−0.045	−0.272	−0.212	−0.100	−0.469	−0.382
Seat all the way	0.617	0.327	0.317	0.404	0.310	0.222	0.776	0.363	0.798
Stand part way	0.000	0.000	0.000	0.015	0.014	0.001	0.010	0.013	0.000
No timetable no map	−0.105	−0.135	0.000	−0.034	0.000	−0.041	−0.112	−0.164	−0.066
Narrow 4-step entry	−0.074	−0.011	0.000	0.000	0.000	0.000	−0.059	0.000	0.000
Wide 2-step entry	−0.373	−0.193	0.000	0.000	0.000	0.000	−0.277	0.000	0.000
Seat under cover at the bus stop	0.354	0.127	0.254	0.159	0.116	0.066	0.180	0.170	0.132
Very friendly driver	0.000	0.036	0.044	0.031	0.049	0.000	0.000	0.053	0.000
Very clean bus	0.000	0.054	0.000	0.000	0.070	0.025	0.000	0.111	0.000
SQI	−1.313	−0.914	−0.476	−1.839	−1.174	−0.463	−1.501	−1.678	−0.915
SQIunsc	−1.335	−1.820	−0.753	−1.839	−1.614	−1.099	−1.409	−1.565	−1.094
Rank order SC	6	3	2	9	5	1	7	8	4
Rank order unsc	4	8	1	9	7	3	5	6	2
SQI Normalised to base 1	1.526	1.925	2.363	1.000	1.665	2.376	1.338	1.161	1.924

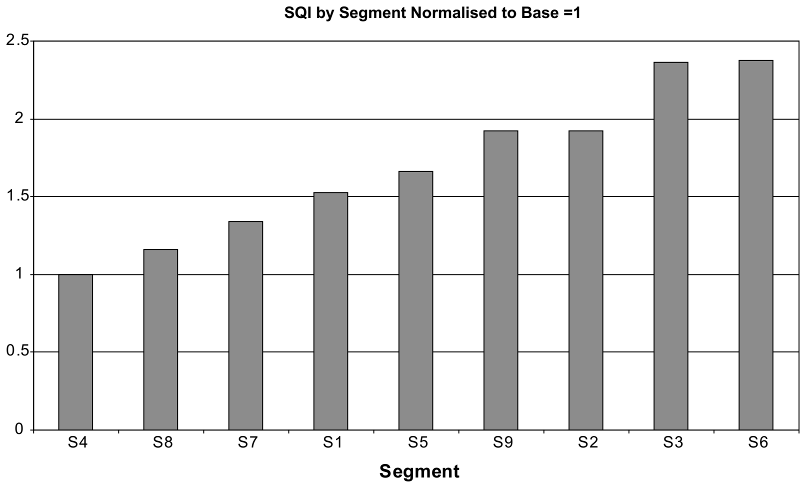

Fig. 2. SQI for each segment normalised to 1.0.

The absolute magnitude of each line in Fig. 3 represents the contribution of an attribute to the overall level of SQI. Each attribute in Fig. 3 is identified by a number for easy tracking. A large positive contribution (above the zero horizontal line) is clearly the preferred outcome, compared to a large negative contribution (below the zero axis). As might be expected, travel time (1) and

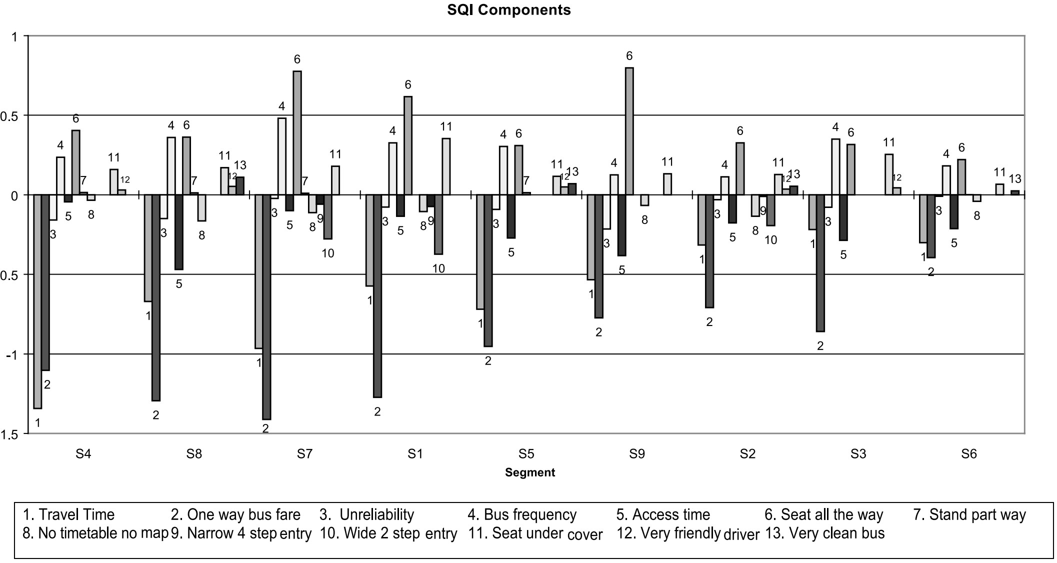

Fig. 3. Decomposition of SQI into its components for each segment.

fare (2) are the greatest sources of negative satisfaction. In comparison, service frequency (4) and getting a seat (6) are the greatest sources of positive satisfaction. Positive or negative satisfactions refer to the passenger's current perception of service level conditioned on the relative importance to the passenger of this service level attribute. Thus the fare level is the greatest contributor to passenger dissatisfaction for all segments except S4 where travel time is a greater contributor (although fare level is still a major concern). It seems clear that reducing fares will be a major contributor to improving SQI, with travel times a close second. Operators might argue that they have limited room to move in adjusting fares and travel times, the former heavily influenced by the regulator and the latter influenced by external factors such as traffic congestion and the general quality of the road environment. Nevertheless it signals a number of issues that operators must address with the other agencies that influence their operating environment.

Looking at the attributes over which the operator has more direct control, having a seat all the way (6) is a source of substantial positive satisfaction, especially for S7, S9 and S1. Access time to bus stop (5) combined with service frequency (4) and service unreliability (3) may be the key drivers of service delivery. All three attributes are substantially under the operator's control and seem to be where the major focus for service improvements should be directed. The provision of infrastructure at the bus stop is a local government obligation in NSW. It appears that passengers in S1 are best served in respect to seat and shelter (11), with passengers in S6 the worst served. Bus cleanliness (13) and driver friendliness (12) have limited relevance to SQI across all segments. In future studies one may reconsider the need for such attributes. One might speculate that these attributes become insignificant in contrast to the fundamental attributes of time, fare, unreliability, comfort (i.e. getting a seat) and service frequency.

The access conditions of the bus (i.e., steps and width) have a significant negative influence for some segments (i.e., S1, S2, S7) with "wide entry with two steps" relative to "wide entry with no

Table 8
Ranking of the 13 attributes in the SQI

Attribute	Segment								
	S1	S2	S3	S4	S5	S6	S7	S8	S9
Positive attributes									
Seat all the way	1	1	2	1	1	1	1	1	1
Bus frequency	3	3	1	2	2	2	2	2	3
Seat under cover at the bus stop	2	2	3	3	3	3	3	3	2
Very clean bus	–	4	–	–	4	4	–	4	–
Very friendly driver	–	5	4	4	5	–	–	5	–
Stand part way	–	–	–	5	6	5	4	6	–
Negative attributes									
Narrow 4-step entry	4	6	–	–	–	–	6	–	–
Unreliability	5	7	5	8	7	6	5	7	5
No timetable, no map at bus stop	6	8	–	6	–	7	8	8	4
Wide 2-step entry	8	10	–	–	–	–	9	–	–
Access Time to bus stop	7	9	7	7	8	8	7	9	6
Travel time	9	11	6	10	9	9	10	10	7
One-way bus fare	10	12	8	9	10	10	11	11	8

steps" having the greatest negative impact. There is clearly room for improvement here, with the potential to increase satisfaction being sufficient to impact the overall SQI rank order of the segments. This sort of diagnosis should be undertaken by each operator to reveal opportunities for service improvement.

Table 8 shows the rankings of the attributes for each segment, arranged so that those contributing positively are listed first, and those contributing negatively are listed last. Those with the lowest rankings have the strongest positive effect on the SQI, while those with the highest rankings have the strongest negative effect on the SQI. As an example, consider segment S5: "having a seat all the way" and "bus frequency" are the two strongest positive contributors; "standing part way" is the smallest positive contributor, and "unreliability" the smallest negative contributor; "travel time" and "one-way bus fare" are the two largest negative contributors. "Wide 2-step entry" is a greater source of negative satisfaction than "narrow 4-step entry." This is clearly an issue for clarification in further studies.

4.5. Determining SQI at the depot level and in the future

The SQI could be determined for a depot, rather than a geographical service segment, although this would lose the variability between segments, which Figs. 2 and 3 show to be quite large. This would be accomplished by estimating an average, weighted by the number of passengers in the segments. More correctly, the models could be re-estimated for each depot using pooled data for the depot instead of for the segments. Similarly, if the desire were to compare segments with each according to depot, this would also require re-estimating the models without the inclusion of any data from another operator (i.e., the segments of one depot in the first case, and those of the other in the second case in this experiment).

If one operator decides to re-survey the same segments at some point in the future and wishes to know if there are changes in the SQI, it would be advisable first to re-estimate the weights for the SQI based on that operator's segments alone, and then apply the new survey attribute levels with the new weights obtained in re-estimation. It would also be possible to apply the new survey attribute levels to the existing weights, but the results are more clearly correct if re-estimated weights are used.

5. Conclusions

This study has progressed the development of SQI at a more detailed level within an organization than the previous pilot study. In addition we have developed and implemented a more rigorous way of identifying the importance weights to attach to statistically significant attributes that recognises the differences in scale between the utility expressions associated with each segment. This is crucial if one is to compare the performance of each segment (i.e., benchmark) meaningfully.

The findings serve a number of purposes. From an operator perspective, they reveal what matters to actual customers and provide some signals as to which attributes need more effort in being marketed to potential patrons. Some of the identified influences on passenger satisfaction are not directly under the control of the bus operator and offer the challenge to influence others (e.g., local government) to contribute to making bus services more attractive. However, recognition of this within the framework of the broader set of influences on passenger satisfaction is very important.

Acknowledgements

We acknowledge the support of the State Transit Authority of NSW and Busways in funding the research presented. The detailed comments of Rhonda Daniels and Stephen Rowe are much appreciated. The advice of Ed Morlok and three referees is also appreciated.

References

Cunningham, L.F., Young, C., Lee, M., 1997. Developing customer-based measures of overall transportation service quality in Colorado: Quantitative and qualitative approaches. Journal of Public Transportation 1 (4), 1–22.

Fielding, G.J., Babitsky, T.J., Brenner, M.E., 1985. Performance evaluation for bus transit. In: Hensher, D.A. (Ed.), Competition and Ownership of Public Transit. A Special Issue of Transportation Research, vol. 19 (1), pp. 73–82.

Hensher, D.A., 1991. Hierarchical stated response designs and estimation in the context of bus use preferences. Logistics and Transportation Reviews 26 (4), 299–323.

Hensher, D.A., 1994. Stated preference analysis of travel choices: the state of the practice. Transportation 21 (2), 107–133.

Hensher, D.A., Bradley, M., 1993. Using stated response data to enrich revealed preference discrete choice models. Marketing Letters 4 (2), 139–152.

Hensher, D.A., Prioni, P., 2002. A service quality index for area-wide contract performance assessment regime. Journal of Transport Economics and Policy (due about mid 2002).

Hensher, D.A., Louviere, J.J., Swait, J., 1999. Combining sources of preference data. Journal of Econometrics 89, 197–221.

Louviere, J.J., Hensher, D.A., Swait, J., 2000. Stated Choice Methods: Analysis and Applications in Marketing, Transportation and Environmental Valuation. Cambridge University Press, Cambridge.

Prioni, P., Hensher, D.A., 2000. Measuring service quality in scheduled bus services. Journal of Public Transport 3 (2), 51–74.

Stopher, P.R., 1998. A review of separate and joint strategies for the use of data on revealed and stated choices. Transportation 25 (2), 187–205.

Swanson, J., Ampt, L., Jones, P., 1997. Measuring bus passenger preferences. Traffic Engineering and Control (June), 330–336.