This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL
REPORT

Value of Reliability of Travellers on the Paris Suburban Railway Network

Technical Report on the Data Analysis

Marco Kouwnhoven, Sebastian Caussade, Eric Kroes

Prepared for the Syndicat des Transports d'Ile de France

RAND EUROPE

The research described in this report was prepared for the Syndicat des Transports d'Ilse de France.

Published 2006 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213
Newtonweg 1, 2333 CP Leiden, The Netherlands
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
Uhlandstraße 14, 10623 Berlin, Germany
RAND URL: http://www.rand.org/
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

The punctuality of trains is an important service characteristic of the public transport system. Trains that suffer from frequent and severe delays cause substantial nuisance to their users and reduce the perceived utility of the system. In other words: irregularity costs money. Therefore rail authorities and train operators look at ways in which they can improve punctuality. In order to appraise the benefits of such measures in relation to the costs associated with them, perceived "values of punctuality" are needed: how much is it worth to a rail traveller to improve punctuality by an amount x?

The objective of the project "*Enquêtes qualitatives et quantitatives visant à évaluer les projets de fiabilisation des radiale ferrées en Ile –de-France*" is to determine a robust method to appraise *a-priori* the monetary benefits of different possible measures to improve regularity. More specifically the use of this methodology is to guide decisions concerning specific projects aiming to make the radial rail lines operate more in line with their published timetables.

The client of this project is the Syndicat des Transports d'Ile de France, based in Paris. The project is carried out by a consortium consisting of STRATEC S.A. (Brussels, main contractor) and RAND Europe (Leiden). The fieldwork and qualitative research has been carried out by Catherine Delannoy and Associates (Paris).

This technical report presents the estimation results of the models that have been calibrated to obtain the key coefficients needed to determine the values-of-reliability. These models have been estimated using a large-scale Stated Preference data set that was collected specifically for this project. Earlier technical memoranda have covered:

1. The characterisation of punctuality
2. A literature review concerning the valuation and quantification of public transport punctuality
3. A statistical analysis of relevant radial rail lines in Paris
4. Qualitative research into the impact of reliability to travellers
5. The specification of the Stated Preference survey designed to measure the valuations.

This document is intended for a technical audience with expert knowledge of discrete choice model estimation using Stated Preference data. A report targeted to a wider, non-specialist audience covering the entire project will be produced separately.

For more information about RAND Europe or this document, please contact Eric Kroes.

Eric Kroes
RAND Europe
Newtonweg 1
2333 CP LEIDEN
The Netherlands
+31 71 524 51 51
reinfo@rand.org

Contents

Summary

This technical report presents the estimation results of the models that have been calibrated to obtain the key coefficients needed to determine the values-of-reliability. These estimation results are needed for the development of a robust method to appraise *a-priori* the monetary benefits of different possible measures to improve regularity of the Paris suburban train network. More specifically the use of this methodology is to guide decisions concerning specific projects aiming to make the radial rail lines operate more in line with their published timetables.

The models have been estimated using a large-scale Stated Preference (SP) data set that was collected specifically for this project. In the SP experiment the respondents had to choose between different service alternatives. Each alternative was described by the travel time, the frequency of short delays (delays between 5 and 15 minutes), the frequency of long delays (delays of more than 15 minutes), the comfort level and the level of information on delays. Each respondent had to make 18 choices between two service alternatives. One choice was added as a check question to examine whether the respondent had understood the SP experiment and was using “common sense” when answering the questions.

More than 2000 travellers were recruited to participate in the SP experiment. Each recruited person received a paper questionnaire with the choice pairs. Their choices were recorded during a telephone interview. From those recruited, about 1300 successful interviews were conducted. Of these, about 30 were excluded from the analysis. This was done either because their reported travel time was outside the agreed upon range, or there was some uncertainty whether the recruited person and the person making the choices were the same person, or they did not really trade between the alternatives (this is: they always choose the left alternative, or vice versa). 1273 respondents were remaining. At a later stage it was decided that the 134 respondents who failed to give the “logical” answer to the check question were excluded.

All estimated coefficients in the final model had intuitive signs and sizes. It was discovered that the extra disutility corresponding to each step in the level of delays decreases as the number of delays increases. The disutility of the first delayed train out of 20 is high, but the extra disutility of the 8^{th} delayed train is much less. Furthermore, the analysis showed that the value of time for people commuting or travelling for education purposes is higher than for people travelling for other purposes. In addition, people travelling for travail/etudes purpose dislike short delays slightly more than people travelling for other purposes. On lines with a high regularity, having a seat is mainly valued for longer trip lengths; on lines with a bad regularity, having a seat is always valued highly, disregarding

the length of the travel. Finally, we found that information about the cause of a delay is always valued highly for journeys towards Paris, but for journeys from Paris this information is more valued if the perceived number of short delays is higher.

The model that has been developed here will be used in the application that is part of the next phase of the project and a separate report covering the entire project will follow.

CHAPTER 1 Design of the Stated Preference Experiment

The design of the Stated Preference (SP) Experiment has previously been discussed in documents concerning that phase of the project [1]. After the pilot results were discussed with the client, it was agreed that some changes should be made to the initial design [2]. In this chapter we briefly outline the final design of the SP experiment used for the main surveys.

Both departing and arriving passengers were recruited on platforms of several train stations in Paris. They are asked about their travel time, the reason for their travel and the level of delay that they perceived. Appendix A lists the lines on which the final set of 1273 respondents were travelling, and also lists their responses to the complete set of recruitment questions.

Based on their responses during the recruitment a questionnaire specific to each respondent was created, printed and sent to them. The advantages of generating individual-specific questionnaires were the increased variation in choices and levels presented to the respondents, which resulted in increased model efficiency. The questionnaire started with a number of Stated Preference choices that had to be made. After this some background questions were asked. The responses to these background questions are given in Appendix B. A few days after this questionnaire had been sent, the respondent was contacted by telephone and his responses were recorded.

The final SP experiment consisted of seven games, each with two or three choices to be made (see Table 1-1). Game 1 was not a real game, but was used to get the recruitment information in a good format on the screen and on the printed questionnaires. Game 8 consisted of only one (dominant) question and was used to check whether the respondent had understood the experiment.

Each choice was between two alternative train services. These services were described by four or five attributes. Table 1-1 gives an overview of the attributes that were varied in each of the games. Each variable attribute could have three levels: a low level, a base level and a high level. Figure 1-1 shows an example from game 7, in which the travel time, the amount of short delay and the information level are varying.

The base level of the first attribute, travel time, was always equal to the travel time reported by the respondent during the recruitment. A random draw determined whether the lower level was either 10%, 20% or 30% below the base level (5% or 10% in cases where the travel time was above 62 minutes). A second draw determined the high level. Table 1-2

Game	Numéro de la question	Niveau de temps de trajet	Niveau des retards de 5 à 15 mn.	Niveau des retards de 15 mn. à 1 h.	Niveau de confort	Niveau d'information
1	*0*	*NOT A REAL GAME : USED TO SHOW RECRUITMENT INFORMATION*				
2	*1-2*	**variable**	**variable**	niveau de base	niveau de base	non proposé
3	*3-4*	niveau de base	**variable**	**variable**	niveau de base	non proposé
4	*5-6*	**variable**	niveau de base	**variable**	niveau de base	non proposé
5	*7-9*	**variable**	**variable**	niveau de base	**variable**	non proposé
6	*10-12*	**variable**	niveau de base	**variable**	**variable**	non proposé
7	*13-15*	**variable**	**variable**	niveau de base	niveau de base	**variable**
8	*16*	**fixed**	niveau de base	**fixed**	**fixed**	niveau de base
9	*17-19*	**variable**	niveau de base	**variable**	niveau de base	**variable**

Table 1-1 Design of SP experiment

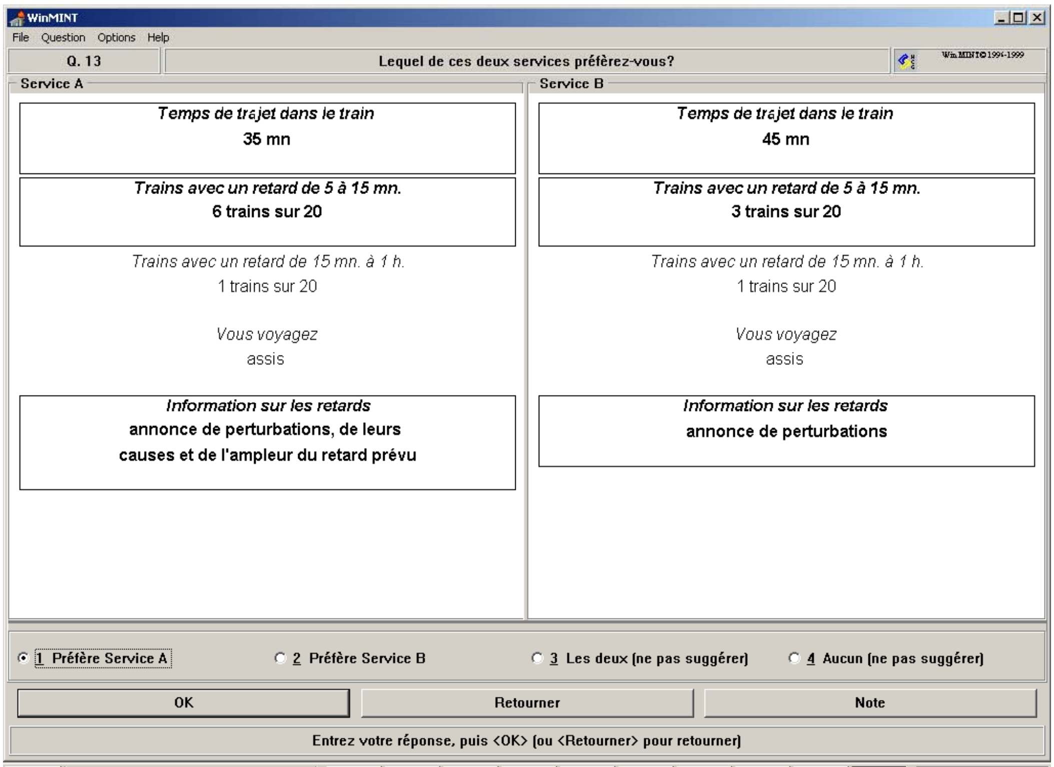

Figure 1-1 Example of a choice (screen as seen by the interviewer)

gives an overview of all possible low, base and high level combinations of the travel time attribute. Note that if the base level was a multiple of 5 minutes, the low and high levels were usually also a multiple of 5 minutes.

Moins que le niveau de base				Niveau de base	Plus que le niveau de base			
-30%	-20%	-10%	-5%	Temps de trajet dans le train	+5%	+10%	+20%	+30%
7	8	8	-	10	-	12	12	13
8	9	9	-	11	-	13	13	14
8	9	10	-	12	-	14	15	16
9	10	11	-	13	-	15	16	17
10	11	12	-	14	-	16	17	18
10	12	13	-	15	-	17	18	20
11	13	14	-	16	-	18	19	21
12	14	15	-	17	-	19	20	22
13	14	16	-	18	-	20	22	23
13	15	17	-	19	-	21	23	25
15	16	18	-	20	-	22	24	25
15	17	19	-	21	-	23	25	27
15	18	20	-	22	-	24	26	29
16	18	21	-	23	-	25	28	30
17	19	22	-	24	-	26	29	31
15	20	22	-	25	-	28	30	35
18	21	23	-	26	-	29	31	34
19	22	24	-	27	-	30	32	35
20	22	25	-	28	-	31	34	36
20	23	26	-	29	-	32	35	38
20	25	27	-	30	-	33	35	40
22	25	28	-	31	-	34	37	40
22	26	29	-	32	-	35	38	42
23	27	30	-	33	-	36	39	43
24	27	31	-	34	-	37	41	44
25	28	30	-	35 – 37	-	40	42	45
25	30	35	-	38 – 42	-	45	50	55
:	:	:	:	:	:	:	:	:
45	50	55	-	58 – 62	-	65	70	75
-	-	55	60	62 – 67	70	75	-	-
-	-	60	65	68 – 72	75	80	-	-
		:	:	:	:	:		
-	-	80	85	88 – 90	95	100	-	-

Table 1-2 Overview of the low, base and high level of the travel time attribute (in minutes)

Moins que le niveau de base	Niveau de base	Plus que le niveau de base
0 sur 20 trains	1 sur 20 trains	2 sur 20 trains
1 sur 20 trains	2 sur 20 trains	4 sur 20 trains
2 sur 20 trains	3 sur 20 trains	6 sur 20 trains
2 sur 20 trains	4 sur 20 trains	8 sur 20 trains
3 sur 20 trains	5 sur 20 trains	10 sur 20 trains
3 sur 20 trains	6 sur 20 trains	12 sur 20 trains
4 sur 20 trains	7 sur 20 trains	12 sur 20 trains
4 sur 20 trains	8 sur 20 trains	12 sur 20 trains

Table 1-3 Overview of the low, base and high level of the short delay attribute (= delays between 5 and 15 minutes)

Moins que le niveau de base	Niveau de base	Plus que le niveau de base
0 sur 20 trains	1 sur 20 trains	2 sur 20 trains
1 sur 20 trains	2 sur 20 trains	4 sur 20 trains
2 sur 20 trains	3 sur 20 trains	6 sur 20 trains
2 sur 20 trains	4 sur 20 trains	8 sur 20 trains
3 sur 20 trains	5 sur 20 trains	8 sur 20 trains

Table 1-4 Overview of the low, base and high level of the long delay attribute (= delays of more than 15 minutes)

The second attribute was the level of short delays. These were delays between 5 and 15 minutes. The level of delay was specified as the number of trains out of 20 that experienced such a short delay. The base level was equal to the level of delays reported during the recruitment, with a minimum of 1 train out of 20 and a maximum of 8 trains out of 20. The low level was about half of the base level. The high level was twice the base level, with a maximum of 12 trains out of 20 (see below). Table 1-3 gives an overview of all possible combinations of levels.

The third attribute was the level of long delays. These were delays of more than 15 minutes. The base level was equal to the level of delays reported during the recruitment, with a minimum of 1 train out of 20 and a maximum of 5 trains out of 20. The low level was about half of the base level. The high level was twice the base level, with a maximum of 8 trains out of 20. The frequency of short and long delays both were presented in the same experiment and therefore their maximum levels delayed had to be such that the total number of trains that were delayed was never above 20. Table 1-4 gives an overview of all possible combinations of levels for the long delay attribute.

The fourth attribute was the level of comfort. The base level was the level reported during the recruitment. In those games in which this attribute was varied, it could have had three levels:

- “assis” (seated)
- “debout” (standing)
- “debout et serré” (standing in a crowded environment)

If the reported level was “toujours debout et serré”, the first level (“assis”) was excluded.

The last attribute was the level of information about delays. This attribute was always varied in those games where it was presented to the respondent, so there was no base level defined. It could have three levels:

- “annonce de perturbations“ (announcement of a disturbance)
- “annonce de perturbations et de leurs causes” (announcement of a disturbance and its cause)
- “annonce de perturbations, de leurs causes et de l'ampleur du retard prévu” (announcement of a disturbance, its cause and the expected amount of delay)

The alternatives shown to the respondents were randomly drawn from all possible combinations. Those combinations in which one alternative was clearly dominating the other alternative (since all attributes were equal or were expected to be valued higher) were excluded. Only in question 16 was one alternative dominating the other. This question was used as a control question to check whether the respondent had understood the SP experiment and was using “common sense” when answering the questions.

Each respondent had the choice to select either alternative A or alternative B. If they indicated to the interviewer that he could not make a choice because they valued them equally, or they thought that neither alternative was acceptable, the interviewer recorded “both” or “none”.

CHAPTER 2 Data Collection

Both the recruitment and the interviews were done by the fieldwork bureau *Catherine Delannoy & Associés – Nexus*. The first two sections of this chapter discuss the recruitment and the data collection. In the third section a problem that occurred during the interviews is explained. An interviewer not following the protocol for opening and closing the respondent files caused a substantial number of interviews to be rejected. Because of this an extra round of interviews were needed. This is discussed in the last section.

2.1 Recruitment

The recruitment took place between May 24th and June 11th 2004 (Monday to Friday only). Both departing and arriving passengers were recruited on platforms of several train stations in Paris. This worked well for the departing travellers, but was less effective for arriving ones. For this reason it was decided as of June 4th 2004 to conduct the recruitments on the trains. In total 2086 persons were recruited.

Between June 17th and 28th the first set of questionnaires were sent to the selected respondents. All questionnaires were accompanied by a letter, in which the respondent was thanked for their cooperation and provided with an explanation of the interview and the SP experiment.

2.2 Interviews

Interviews were performed by a team of 7 interviewers. They tried to call the respondents on weekdays between 16:00 and 21:00h and on Saturdays between 10:00 and 20:00h. Interviews started on Wednesday June 23rd and continued for eleven days; during this period a total of 1211 interviews were completed. From the 2086 recruitments, 131 incorrect telephone numbers were recorded, 75 refused to participate, 30 were on holiday and 9 persons had not received their questionnaire by post. 630 recruited persons could not be successfully contacted within the interviewing timeframe.

2.3 Problem of Corrupted Files

On July 7th CDA delivered 1211 data files with results from the interviews to RAND Europe. On delivery it was discovered that 235 of these files were corrupted. In almost all cases the corruption was caused by an interviewer not following the correct protocol for opening and closing the respondent files.

In the CATI[1] program (WinMINT), data from two sources needed to be combined: information from the recruitment and information from the printing of the questionnaires that were sent to the respondent. During this printing the customised choice pairs that were to be presented to the respondent were selected. When the CATI protocol was not followed, it was possible to have a situation where the choice pairs on the printed questionnaire in front of the respondent did not match the choice pairs on the screen of the interviewer, though this might only show in the second, third or even a later choice pair.

In those cases it is not unambiguously clear whether the respondent made his choice between the alternatives shown on his paper or between the alternatives read to him by the interviewer. For this reason these respondents were rejected from further analysis and their files were marked "corrupt".

As a result of the deviation from the protocol, it was also possible for a WINMINT data file that should have contained the responses for one respondent only to contain the choices made by two or more respondents. In some specific cases the choices made by one of these respondents could be unambiguously recovered. This process is explained in section 3.1.

Because some respondent files contained choices from multiple respondents, other respondent files did not contain any choice information at all. These files were marked as "corrupt" as well and were not recoverable.

In a few cases it was clear from a respondent file that an attempt had been made to contact another respondent (while the wrong respondent file was open). Fortunately, this other respondent was not at home, so only the (incorrect) recruitment information was copied into the file of the original respondent. Once, this original respondent was successfully contacted, the correct recruitment information was rewritten to the file. The analysis software only uses the last written information, so these files were *not* marked as "corrupt" and were handled in the normal way.

In a few other cases the respondent file showed unexpected results: the same respondent seemed to be contacted twice and in both cases different choices were recorded. Those cases were marked "corrupt" and were not recoverable.

2.4 Additional Interviews

After the problem with the corrupted files was discovered, a second phase of the interviews was started. In this phase recruited persons that were not yet contacted during the first phase were called. In addition, in the week following July 12th, copies of the original questionnaires were sent again to respondents for which the first CATI resulted in a

[1] Computer Assisted Telephone Interview

Data delivery number (Wave)	Date	Number of files	Number of corrupt files	of which are recoverable	Number of uncorrupted files after recovery
1	7 July	1211	235	22	998
2-4*	19-27 July	121	9	2	114
5	29 July	100	3	0	97
6	31 July	28			28
7	9 August	29	2	0	27
8	12 August	9			9
9	18 August	10			10
10	25 August	8	1	0	7
11	3 September	13			13
Total		**1529****	**250**	**24**	**1303**

* overlapping deliveries
** this includes identical respondents for which a corrupt file was delivered during the first phase of the CATI and an uncorrupted file during a later phase

Table 2-1 Overview of the data file deliveries

corrupted file. These respondents were contacted and were told that something had gone wrong with the recording of their choices and they were asked to make their choices again.

During July and August, CDA delivered 317 additional data files, 14 of which suffered from similar problems as before and turned out to be corrupt. An overview of the data file deliveries is given in Table 2-1. All data files are grouped into “waves” corresponding to the index number of the delivered set of files. The interview phase ended early September. At that time only a small number of the originally recruited persons had not been contacted, although this contact had been attempted on many occasions.

CHAPTER 3 Data Processing

Before the data collected during the interviews could be used to estimate the final model, some data processing was needed. The first step of this processing has already been mentioned in the previous chapter: some of the data files turned out to be corrupt but in a few cases the information in the files could be recovered. This recovery process is explained in detail in the first section of this chapter. The next step in the processing was the inspection of the quality of the data. A few files did not pass the quality tests and were excluded from further analysis. This is explained in the second section of this chapter. The chapter ends with an inspection of the distribution of the final set of respondents over the quota variables and targets set prior to the fieldwork.

3.1 Data Recovery

This section provides an overview of the process in which some of the choice data was recovered from corrupted data files. Files that could be recovered can be divided into three groups:

Type 1: files with choice data from multiple respondents, but ending with the choice data from the correct respondent. These files did not need any manual changes.

- Respondent file 3052 (wave 4). When this file was open, an unsuccessful attempt was made to start an interview with respondents 3068, 3072, 3078, 3086. These persons were probably not at home. Respondent 3118 was reached and his choice data was recorded. Since the choice alternatives on the screen were from respondent 3052, the data from respondent 3118 is ambiguous and cannot be used for further analysis. Two days later (July 22) respondent 3052 was successfully contacted and his choices were recorded. These choices overwrite the previously recorded choices from respondent 3118, so this file can be used in the normal way.
- Similar problem occurred with respondents: 53, 138, 273, 434, 451, 620, 639, 642, 643, 1811, 2046, 2113, 2409, 2579, 2925, 3034, 3260, 3263 (all wave 1)

Type 2: files with choice data from multiple respondents, amongst which the choice data from the correct respondent. The last respondent in the data file is not the correct respondent. These files needed manual changes.

- Respondent file 359 (wave 1). After the choice data of this respondent was successfully stored, a copy of the data file was generated. About a week later (July

2) the data file for respondent 359 was opened again, but now the choice data for respondent 3052 was stored. This data file with choice data for two respondents was manually replaced by the copy that was made after the first (and correct) interview. Note that respondent 3052 was contacted again at a later stage (wave 4), its data file was corrupted, but could be recovered (see above).

- Similar problem occurred with respondent file 9 (wave 2).
- Respondent file 506 (wave 1). After the choice data of this respondent was successfully stored, choice data from respondent 551 was stored in the same file. The file was not closed in between, so no copy was made. All recorded information from respondent 551 was manually deleted from the file.

Type 3: files that were incorrectly closed.

- Respondent file 406 (wave 1). After the respondent file is opened during an unsuccessful attempt to contact the respondent, the data file did not close correctly (for some unknown reason). It could be that the computer was reset, so that the last line (which records the closing date and time) was not written to the file. When the file was reopened again, it contained two successive opening lines, without a closing line. After this the choice data was successfully recorded, however, the data analysis software cannot handle this file because of the two opening lines. One of the opening lines was manually removed.
- Similar problem occurred with respondent file 3189 (wave 1).

3.2 Inspection of the Data Quality: Exclusions

Some of the reasons to exclude respondents from further analysis have already been described in the section on corrupt files. In addition to these, we made the following checks to determine whether a respondent should be included in the response analysis.

1. **Check on travel time**
   It was agreed with the client that the travel time reported by the respondent should be between 10 and 120 minutes. Respondent 3321 reported a travel time of 5 minutes during the recruitment. Respondent 3096 reported a travel time of 45 minutes during the recruitment, but changed this into 8 minutes during the CATI. Respondents 2, 10, 50 changed their travel time to 25 hours during the CATI from 15, 25 and 50 minutes, respectively. They might have meant to change to 25 minutes, but based on their origin and destination stations, this did not make sense in all cases. Therefore, all the respondents mentioned in this paragraph were excluded from further analysis.

2. **Check on gender**
   During the recruitment the recruiter reported the gender of the recruited person. During the CATI interview the interviewer should have checked when starting the interview that the respondent was the same person as was recruited. At the end of the interview the interviewer recorded the gender of the person. We have checked

Data delivery number (Wave)	Number of uncorrupted files	Excluded because of travel time out of range	Excluded because of gender inconsistency	Excluded because of lack of trading	Number of files included in analysis
1	998	4	20	2	972
2-4	114		2		112
5	97				97
6	28		1		27
7	27	1			26
8	9				9
9	10				10
10	7				7
11	13				13
Total	**1303**	**5**	**23**	**2**	**1273**

Table 3-1 Overview of the number of excluded respondents

whether the gender reported during the recruitment was identical to the one reported during the CATI. This was not the case for 30 respondents. CDA has tried to contact these persons again. In 7 cases the respondent confirmed he was interviewed during both the recruitment and the CATI. In those cases it is likely that a mistake was made when reporting the gender during either the recruitment of the CATI. The remaining 23 respondents were excluded from further analysis.

3. **Check on trading**
   Two respondents always choose the left alternative (1) or always choose the right alternative (2) shown on paper. It is possible that these persons have not seriously considered both alternatives and for safety both respondents were excluded from further analysis.

Table 3-1 gives an overview of all the exclusions for each wave of data. 1273 respondents have been used in the analysis phase.

3.3 Distribution over the Quota

Table 3-2 compares the distribution of the respondents over the quota variables and their corresponding targets (between brackets). From this table it can be concluded that the targets are well matched, except for the number of respondents with "other" travel purposes, which is slightly low. The total number of respondents is higher than agreed on. This compensates for segments of respondents with increased unexplained error (= slightly reduced data quality), see section 4.3 and 4.4.

An attempt was also made to get an equal distribution over the travel direction (towards / from Paris), the day of the week, and time-of-day (from 7:00 to 19:00h), see Table 3-3 to Table 3-5. From these tables it can be concluded that the respondents are about equally divided between travellers to and from Paris. The spread over the days is quite reasonable, although there seems to be a lack of respondents recruited on Mondays. In the distribution over time-of-day one can see a preference towards the early morning and late afternoon, consistent with the times that most people travel to and from work. About 10% of the respondents are recruited before 7:00h or after 19:00h. Table 3-6 splits the distribution over time-of-day between travellers towards and from Paris. The morning peak has been defined here from 7:30 to 9:30h. From this table it can be seen that the number of travellers towards and from Paris are both distributed over time-of-day in a similar way in the sample collected.

Frequency + Regularity combination Purpose	Basse Fréquence; Bonne Régularité	Basse Fréquence; Mauvaise Régularité	Haute Fréquence; Bonne Régularité	Haute Fréquence; Mauvaise Régularité	Total
Domicile <-> travail + Domicile <-> études	195 *(200)*	266 *(200)*	195 *(200)*	249 *(200)*	**905** ***(800)***
Autres	92 *(100)*	83 *(100)*	106 *(100)*	87 *(100)*	**368** ***(400)***
Total	**287** ***(300)***	**349** ***(300)***	**301** ***(300)***	**336** ***(300)***	**1273** ***(1200)***

Table 3-2 Distribution of the respondents over the quota variables

	Frequency	Percent
From Paris (departing passenger)	624	49.0
To Paris (arriving passenger)	649	51.0
Total	**1273**	**100.0**

Table 3-3 Distribution of the respondents over their travel direction

	Frequency	Percent
Monday	88	6.9%
Tuesday	239	18.8%
Wednesday	339	26.6%
Thursday	253	19.9%
Friday	353	27.7%
Saturday	0	0.0%
Sunday	0	0.0%
Not recorded	1	0.1%
Total	**1273**	**100.0**

Table 3-4 Distribution of the respondents over the day of the week of their recruitment

	Frequency	Percent
0:00 - 4:59	4	0.3%
5:00 - 6:59	63	4.9%
7:00 - 8:59	194	15.2%
9:00 - 10:59	254	20.0%
11:00 - 12:59	185	14.5%
13:00 - 14:59	120	9.4%
15:00 - 16:59	178	14.0%
17:00 - 18:59	213	16.7%
19:00 - 20:59	51	4.0%
21:00 - 23:59	10	0.8%
Not recorded	1	0.1%
Total	**1273**	**100.0**

Table 3-5 Distribution of the respondents over the time of the day of their recruitment

		DIRECTION		
		Departing passenger	Arriving passenger	Total
0:00 - 7:29		37	64	101
7:30 - 9:29	(Morning peak)	103	125	228
9:30 - 14:59		250	241	491
15:00 - 16:59	(Early evening peak)	93	85	178
17:00 - 18:59	(Late evening peak)	102	111	213
19:00 - 23:59		39	22	61
Not recorded		0	1	1
Total		**624**	**649**	**1273**

Table 3-6 Distribution of the respondents over the time of the day of their recruitment, split by direction of the journey

CHAPTER 4 Model Development

The development of the model estimated from the data of the SP experiment consists of four steps. After an initial model was estimated (section 4.1), tests were run to explore whether certain segments of the data contained different amounts of unexplained error and hence needed a different scaling (sections 4.2, 4.3 and 4.4). After this, tests were run to investigate whether some segments of the data contained significantly different response behaviour (section 4.5). In the final phase of the model development, piecewise-linear coefficients were tried (section 4.6), an error component model was tested (section 4.7) and a jack-knife procedure was performed to estimate the final T-ratio's (section 0).

4.1 First Models

As shown in the previous chapter, 1273 respondents remained in the sample after exclusions 1, 2 and 3 were applied (section 3.2). Each respondent made 18 choices (+ 1 extra choice when answering the check question), resulting in 22,914 observations. The answer "both" was recorded in 12 observations, while "none" given in 17 cases. These 29 observations were excluded from further analysis.

The first model that we tried to estimate included coefficients for all the main attributes that were presented in the SP choices and cross terms between these attributes. These terms and coefficients are explained in Table 4-1.

The results can be found in Table 4-2 (model "FIRST"). All the coefficients on the base terms are significant at a 95% confidence level (T-ratio of 2 or above). A few of the coefficients on the cross terms turn out to be insignificant. However, they were kept in the model, because they may have become more significant in later stages, e.g. when segmentation was applied. Appendix C explains the modelling in more detail. It also elaborates on the model parameters as shown in Table 4-2.

4.2 Comparison of the Different Games

The second step in the model development was to check whether or not the different games contained different information and whether the amount of unexplained error in the games was compatible. To check this we created a utility function that applied a different scale factor to each of the games, see Figure 4-1 for the tree structure. These scale

Coefficient	Multiplied by	Applied only when	Remark
Base terms			
time	time		Time level shown on card
delay10	delay10		Number of trains out of 20 with a delay between 5-15 min.
delay30	delay30		Number of trains out of 20 with a delay.of more than 15 min.
comfort2		Comfort level on card = 2	"debout"
comfort3		Comfort level on card = 3	"debout et serré"
info2		Information level on card = 2	"annonce de perturbations et de leurs causes"
info3		Information level on card = 3	"annonce de perturbations, de leurs causes et de l'ampleur du retard prévu"
Cross terms			
timecomf2	time	Comfort level on card = 2	
timecomf3	time	Comfort level on card = 3	
infody102	delay10	Information level on card = 2	
infody103	delay10	Information level on card = 3	
infody302	delay30	Information level on card = 2	
infody303	delay30	Information level on card = 3	
tvdelay10	(time * delay10)		
tvdelay30	(time * delay30)		
tvinfo2	time	Information level on card = 2	
tvinfo3	time	Information level on card = 3	
dly10com2	delay10	Comfort level on card = 2	
dly10com3	delay10	Comfort level on card = 3	
dly30com2	delay30	Comfort level on card = 2	
dly30com3	delay30	Comfort level on card = 3	

Table 4-1 Explanation of the terms and coefficients used in the model development

factors indicate the relative amount of unexplained error compared to game 2. If this amount were similar between the data sources then a scale factor of 1 would be expected.

The results can be found in Table 4-2 (model "SECOND"). From this, it follows:

- The scale does not deteriorate significantly across the games, implying that there is no fatigue effect across the 19 choices in which all respondents participated.
- Scale 3 is significantly lower than 1, indicating that game 3 has more unexplained error than game 2. (In Game the time and delay30 attributes are varied
- Scale 5 and scale 6 are both significantly different from 1, but they are not significantly different from each other. Both games are similar: time, comfort and either delay10 or delay30 is varied.
- Scales 4, 7 and 9 are not significantly different from 1.

Model	FIRST		SECOND		SECOND-B	
Observations		22885		22885		22885
Final log (L)		-15092.5		-15081.0		-15085.2
D.O.F.		21		27		23
Rho²(0)		0.049		0.049		0.049
Rho²(c)		0.047		0.048		0.048
Base terms	***Coeff***	***T-ratio***	***Coeff***	***T-ratio***	***Coeff***	***T-ratio***
time	-0.0356	(-16.3)	-0.0422	(-9.2)	-0.0408	(-15.2)
delay10	-0.125	(-13.6)	-0.143	(-9.3)	-0.139	(-12.9)
delay30	-0.179	(-10.8)	-0.203	(-7.3)	-0.196	(-10.4)
comfort2	-0.411	(-5.0)	-0.456	(-3.7)	-0.456	(-4.1)
comfort3	-1	(-11.3)	-1.25	(-7.1)	-1.24	(-9.0)
info2	0.248	(2.7)	0.25	(2.2)	0.203	(2.2)
info3	0.669	(7.2)	0.713	(4.9)	0.619	(6.6)
Cross terms						
timecomf2	-0.0091	(-4.5)	-0.0134	(-4.1)	-0.0128	(-4.3)
timecomf3	-0.0096	(-4.3)	-0.0142	(-4.0)	-0.0133	(-4.2)
infody102	0.0122	(1.0)	0.0188	(1.4)	0.0201	(1.6)
infody103	0.0015	(0.1)	0.0059	(0.4)	0.0104	(0.9)
infody302	0.0068	(0.3)	0.0097	(0.4)	0.0135	(0.7)
infody303	-0.0167	(-0.9)	-0.013	(-0.6)	-0.01	(-0.5)
tvdelay10	-5.80E-04	(-2.2)	-9.30E-04	(-2.7)	-8.90E-04	(-2.8)
tvdelay30	-9.10E-04	(-1.9)	-0.0016	(-2.7)	-0.0015	(-2.6)
tvinfo2	-0.0029	(-1.4)	-0.003	(-1.3)	-0.0025	(-1.2)
tvinfo3	-0.0045	(-2.2)	-0.0046	(-1.9)	-0.004	(-1.9)
dly10com2	0.0094	(0.8)	-0.0068	(-0.4)	0.0028	(0.2)
dly10com3	-0.0143	(-1.0)	-0.0544	(-2.0)	-0.0273	(-1.4)
dly30com2	-0.0013	(-0.1)	-8.50E-04	(-0.0)	-0.0091	(-0.3)
dly30com3	-0.0343	(-1.5)	-0.0251	(-0.7)	-0.0511	(-1.6)
Scale factors						
Scale3			0.693	(2.7)	0.713	(2.6)
Scale4			1.09	(0.6)	1	
Scale5			0.633	(4.3)		
Scale6			0.777	(2.3)		
Scale7			0.935	(0.6)	1	
Scale9			0.864	(1.1)	1	
Scale5+6					0.738	(4.0)

Table 4-2 Model parameters and estimated coefficients of the FIRST, SECOND and SECOND-B model. All T-ratios are with respect to 0, except the T-ratios on the scales, which are with respect to 1 (also in other tables).

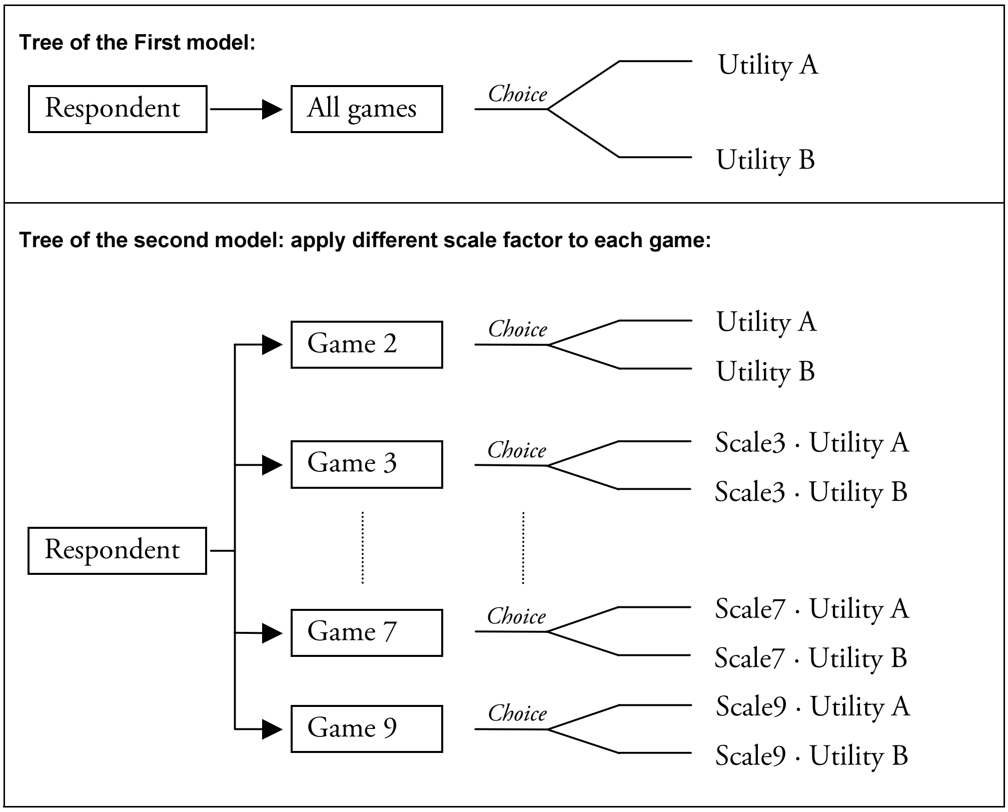

Figure 4-1 Tree structure of the first and second model

Therefore, we made another run, with scales 4, 7 and 9 fixed to 1, scale 5 and scale 6 combined into one scale factor 5+6. This run is called SECOND-B and its coefficients are also displayed in the following table.

Based on a log-likelihood test, we conclude that model SECOND-B is better than the FIRST model. The difference between SECOND and SECOND-B is on the edge of statistical significance (at the 95% confidence level). For reasons of simplicity, we preferred the SECOND-B model and used this as the basis for further development.

4.3 Dominant Question

In the 16th choice pair all the attributes of the second alternative (right hand side in Figure 4-2) were favourable over the attributes of the first alternative (left hand side). This is the so-called dominant question. It would be expected that if a respondent understood the questionnaire and was still paying attention to the differences between the alternatives they would choose the second alternative (service B).

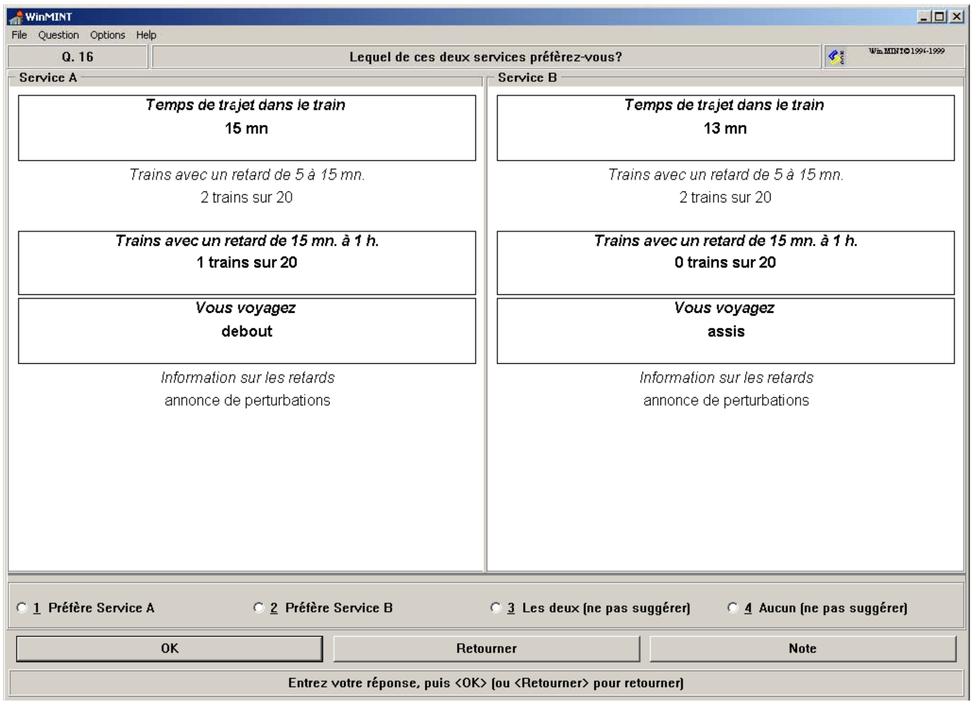

Figure 4-2 WINMINT screen for question 18 (dominant question)

Data delivery number (Wave)	Number of files included in analysis	Fail to pass check on dominant question	Fail to pass check on sum trains delayed
1	972	132	25
2-4	112	1	1
5	97		3
6	27	1	2
7	26		
8	9		
9	10		1
10	7		
11	13		1
Total	**1273**	**134**	**33**

Table 4-3 Overview of the number of respondents not passing checks on the dominant questions and the sum of the numbers of trains delayed

Of the 972 uncorrupted files in the first data delivery (after exclusions 1 – 3) 132 respondents gave the non-intuitive answer to the dominant question (13%). (Non-intuitive includes "both" or "none"). On July 23rd CDA changed their protocol so that persons that did not give "Service B" as their choice for the dominant question, were asked whether or not they were sure. According to CDA respondents did not often change their choice. However, of the 301 respondent files that were delivered after the first data delivery, only 2 respondents did not choose "Service B" (1%), see Table 4-3.

First, we created a model in which an extra scale factor, called *additional*, was applied to the data from interviews after July 23rd. A second scale factor, called *dominant*, was applied to all respondents who gave a non-intuitive answer to the dominant question. The size of the scale factor indicates whether these segment of respondents have a higher amount of unexplained error in their answers, compared to respondents from before July 23rd, that gave the expected answer to the dominant question (i.e. respondents which we believe understood the SP experiment). The results of this model are displayed in Table 4-4 in the column THIRD.

From this table it can be concluded that the amount of unexplained error for the respondents with an unexpected answer to the dominant question is much larger than for the other respondents (this is shown by the very small scale, which compensates for the increased error). The scale factor *dominant* is still significantly different from zero (T-ratio 2.2), so some information is contained in this data segment. To test whether including these respondents improves the model, we have made an extra run (THIRD_B) in which these respondents are excluded. Since these two models (THIRD and THIRD_B) contain a different number of observations, the T-ratios of the coefficients need to be compared to judge which is the better model. As can be seen, there is little difference between the T-ratios in the two models. The Rho^2 value is better for the THIRD-B model.

This check was also repeated again at a later stage of the modelling. At that stage the T-ratios of the model that excluded the respondents with a non-intuitive answer to the dominant question, was slightly better (there were a few more T-ratios that improved by 0.1/0.2 than there were T-ratios that decreased by 0.1). For this reason we decided to exclude these respondents from further analysis.

The scale factor *additional* is clearly different from 1. This means that answers from respondents in the later phases have a higher unexplained error level than the respondents from the first phase (until 23rd of July). This could have several causes:

- It might be that we have sampled a different segment of people. Many of the respondents in the later waves had previously been tried to be contacted in earlier stages, but could not be reached. However, they were eventually reached during the summer holiday period.
- For these people there was quite some time between the recruitment and the actual interview. Their perceptions or their memory of their journey might have changed in this time.
- It might be that during the summer holiday people perceive travel time, delays, comfort, and information differently than during the period just before the holidays.

Model	THIRD		THIRD-B		FOURTH	
Observations		22885		20481		20481
Final log (L)		-15004.2		-13344.2		-13342.3
D.O.F.		25		24		24
Rho²(0)		0.054		0.060		0.060
Rho²(c)		0.052		0.059		0.059
Base terms	***Coeff***	***T-ratio***	***Coeff***	***T-ratio***	***Coeff***	***T-ratio***
time	-0.0483	(-16.0)	-0.0486	(-16.1)	-0.0492	(-16.1)
delay10	-0.168	(-13.6)	-0.169	(-13.7)	-0.173	(-13.7)
delay30	-0.249	(-11.6)	-0.252	(-11.7)	-0.256	(-11.6)
comfort2	-0.452	(-3.6)	-0.449	(-3.6)	-0.449	(-3.5)
comfort3	-1.43	(-9.3)	-1.44	(-9.3)	-1.45	(-9.2)
info2	0.226	(2.2)	0.222	(2.1)	0.212	(2.0)
info3	0.709	(6.8)	0.707	(6.8)	0.698	(6.6)
Cross terms						
timecomf2	-0.0157	(-4.6)	-0.0157	(-4.6)	-0.0161	(-4.6)
timecomf3	-0.0153	(-4.2)	-0.0153	(-4.2)	-0.0158	(-4.2)
infody102	0.0237	(1.7)	0.024	(1.7)	0.0264	(1.8)
infody103	0.0057	(0.4)	0.0044	(0.3)	0.0056	(0.4)
infody302	0.0225	(1.0)	0.0237	(1.0)	0.0229	(1.0)
infody303	0.0018	(0.1)	0.0041	(0.2)	0.009	(0.4)
tvdelay10	-8.50E-04	(-2.4)	-8.10E-04	(-2.3)	-8.40E-04	(-2.4)
tvdelay30	-0.0015	(-2.4)	-0.0015	(-2.3)	-0.0016	(-2.5)
tvinfo2	-0.003	(-1.3)	-0.0029	(-1.3)	-0.0027	(-1.2)
tvinfo3	-0.0036	(-1.5)	-0.0034	(-1.4)	-0.0033	(-1.4)
dly10com2	-0.0119	(-0.6)	-0.0132	(-0.7)	-0.0123	(-0.7)
dly10com3	-0.0296	(-1.3)	-0.0292	(-1.3)	-0.0293	(-1.2)
dly30com2	-0.0116	(-0.4)	-0.011	(-0.4)	-0.0116	(-0.4)
dly30com3	-0.081	(-2.1)	-0.0811	(-2.1)	-0.0832	(-2.1)
Scale factors						
Scale3	0.747	(2.4)	0.755	(2.3)	0.75	(2.4)
Scale5+6	0.729	(4.3)	0.725	(4.4)	0.717	(4.6)
Additional	0.758	(4.1)	0.758	(4.1)	0.758	(4.1)
Dominant	0.139	(13.6)				
SumTrains					0.734	(2.1)

Table 4-4 Model parameters and estimated coefficients of the THIRD, THIRD-B and FOURTH model

- Some of the respondents in this additional stage were already contacted during the first stage, but their choices were not stored correctly (corrupt files). It might be that these people responded differently.
- It might be that a similar percentage of respondents answered the dominant question in a non-intuitive way, but they changed their choice when the interviewer asked whether they were sure. As seen before, these respondents have a very high level of unexplained error. They are mixed with "normal" respondents, so the amount of error in the mixed sample is increased.

4.4 Sum of Number of Trains Delayed out of 20

During the recruitment the respondents were asked for the number of trains out of 20 that they believe are delayed between 5 and 15 minutes and for the number of trains out of 20 that are delayed for more than 15 minutes. Obviously, the sum of these two numbers cannot exceed 20 out of 20. However, 33 persons provided numbers of trains delayed that resulted in a sum that exceeded 20, see Table 4-3.

It might be that these people did not understand the concept of "number of trains out of 20" correctly. It might also be that the respondents wanted to give a signal that they were not satisfied with the present reliability. All this might cause extra unexplained error in their choices. To check this hypothesis, a test was performed where an extra scale factor *SumTrains* was applied to these respondents. The result can be found in Table 4-4 in the column on model FOURTH.

The resulting scale factor is just significantly different from 1 (at a 95% confidence level the T-ratio for this deviation is 2.1). So, there is some extra unexplained error in the choices of these respondents, however, this is not a very strong effect. For this reason, these observations were kept in the model, and the extra scale factor was retained in the following models.

4.5 Segmentations

Tests were performed to examine whether the model coefficients varied according to some predetermined exogenous factors (e.g. journey purpose, frequency). For this, we have made separate models for different data segments. Tests were performed for each of the following quota variables:

- **Purpose (domicile <-> travail/études versus other)**
  The separate runs for the two purpose segments revealed that only the *time* and *delay10* coefficients might be different for each purpose. A single model with separate coefficients for both time and delay10 for each purpose showed that the difference between these coefficients was indeed significant, with commuters valuing travel time more negatively than those travelling for other purposes.

- **Frequency (basse fréquence versus haute fréquence)**
  Based on the separate runs, we tried different coefficients for *time, delay10, delay30, comfort2, comfort3, info2* and *info3* for two different frequency segments. None of these coefficients turned out to be significantly different between the two frequency segments.

- **Regularity (bonne regularité versus mauvaise regularité)**
  Based on the separate runs, we tried different coefficients for *time, delay10, delay30, comfort2, comfort3, info2, info3, timecomf2* and *timecomf3.* The segment-specific coefficients for *delay10, delay30, comfort2,* and *timecomf2* turned out to be mutually significantly different.

- **Direction (travelling towards Paris versus from Paris)**
  Based on the separate runs, we tried different coefficients for *time, delay10, delay30, comfort2, comfort3, info2, info3, infody102* and *infody103.* The segment-specific coefficients for *delay30, info2, infody102* and *infody103* turned out to be mutually significantly different.

These segmentations were each undertaken individually. To consolidate these we constructed one single model with all these segmentations. This revealed that that many of the coefficients that were segmented on the basis of the analysis described above, could not pass the log-likelihood test in the combined model. The segmentations that remained were:

- The *time* coefficient is significantly different (1.5 times more negative) for journeys made for the domicile <-> travail/etudes purpose and journeys made for other purposes. The value of time for people commuting or travelling for education purposes is higher than for people travelling for other purposes.

- The *delay10* coefficient (for delays between 5 and 15 minutes) is significantly bigger (1.3 times more negative) for journeys made for the domicile <-> travail/etudes purpose than for journeys made for other purposes. This means that people dislike arriving late for/from their work / study more than arriving late for/from other purpose destinations. We did not see a similar effect on the *delay30* coefficient.

- The *delay30* coefficient (for delays of more than 15 minutes) is significantly bigger (1.3 times more negative) for journeys made with lines with a good regularity than for journeys made with lines with a bad regularity.

- The *comfort2* coefficient is significantly bigger (3.5 times more negative) for journeys made on lines with a poor regularity than for journeys made on lines with a good regularity. If there is a higher probability of a delay, people value having a seat more than for lines with a lower probability of a delay. This effect is partly compensated by the fact that the *timecomf2* coefficient is higher on lines with a high regularity. The conclusion here is that on lines with a high regularity, having a seat is mainly valued for longer trip lengths; on lines with a bad regularity, having a seat is always valued highly, regardless of the length of the journey.

Model	**FIFTH**		
Observations		20481	
Final log (L)		-13324.3	
D.O.F.		23	
Rho²(0)		0.061	
Rho²(c)		0.060	
Base terms	***Coeff***	***T-ratio***	***Applied only when***
time	-0.0548	(-16.7)	Purpose = commute / education
	-0.0375	(-8.4)	Purpose = other
delay10	-0.186	(-14.2)	Purpose = commute / education
	-0.150	(-9.9)	Purpose = other
delay30	-0.283	(-11.3)	Regularity = good
	-0.220	(-9.5)	Regularity = bad
comfort2	-0.238	(-1.7)	Regularity = good
	-0.828	(-5.6)	Regularity = bad
comfort3	-1.52	(-10.2)	
info2	0.0041	(0.0)	Direction = from Paris
	0.291	(4.8)	Direction = towards Paris
info3	0.631	(13.2)	
Cross terms			
timecomf2	-0.0217	(-4.6)	Regularity = good
	-0.009	(-2.3)	Regularity = bad
timecomf3	-0.0152	(-4.2)	
infody102	0.0610	(3.4)	Direction = from Paris
	0		Direction = towards Paris
tvdelay10	-7.2E-04	(-2.0)	
tvdelay30	-0.0019	(-2.8)	
dly30com3	-0.0903	(-2.6)	
Scale factors			
Scale3	0.737	(2.6)	
Scale5+6	0.73	(4.6)	
Additional	0.758	(4.1)	
SumTrains	0.764	(1.8)	

Table 4-5 Model parameters and estimated coefficients of the FIFTH model

- The *info2* coefficient (announcing the cause of a delay) is significant for journeys made towards Paris. This coefficient is equal to zero for journeys made from Paris. However, the *infody102* cross term coefficient (infolevel2, delay10) is only significant for journeys from Paris. This means that the information about the cause of a delay is always valued highly for travel towards Paris, but for travel from Paris it is valued more highly if the perceived number of short delays is higher.

Table 4-5 gives the results of this model (FIFTH). All insignificant cross terms have been removed from the model.

In addition to the segmentations based on the quota variables discussed above, we also checked whether the respondents needed to be segmented based on their answers to other recruitment and background questions. This was done with tables, which provide the predicted and actual distribution of choices (based on the answers of one background of recruitment question). If the answers to this question are correlated to the choice behaviour, a comparison of the predicted and actual distribution of choices reveals systematic differences. We checked segmentations based on the following questions: perceived comfort level, perceived information level, whether the respondent's arrival time is flexible, whether a connecting trip had to be made, what kind of ticket was used, how often this trip was made, whether going by car was an alternative, the age of the respondent and the time-of-day of the recruitment. None of these had an impact on the valuation of travel time, delay level (short or long), comfort level or information level, so no further segmentations were possible to explain systematic variations.

4.6 Piece-wise linear coefficients

The models presented have all contained a linear term for the journey time. A test was performed to examine whether a logarithmic journey time specification led to any improvements in model fit. In practice this was found to lead to a deterioration in the model fit. In addition, tests were undertaken to examine whether there may be benefit from having separate time coefficients for the travel times less than 20 minutes, between 20 and 35 minutes and above 35 minutes. It turned out that these three coefficients were not significantly different, so the best model was obtained with a single linear time coefficient.

Tests were also undertaken to see whether benefits could be obtained from making the two delay coefficients piece-wise linear. All possible delay levels were split into three intervals: "1-3 out of 20", "4-6 out of 20" and "7 or more out of 20". Separate coefficients were applied to each of these intervals. The utility terms corresponding to these intervals were added in a cumulative way to ensure continuity. This means that if a certain alternative was described by a delay level of "5 trains out of 20", the corresponding utility had a component for 1-3 journeys delayed, plus a component for 4-6 journeys delayed

Model	**SIXTH**		
Observations		20481	
Final log (L)		-13266.7	
D.O.F.		25	
Rho²(0)		0.065	
Rho²(c)		0.064	
Base terms	***Coeff***	***T-ratio***	***Applied only when***
time	-0.065	(-18.2)	Purpose = commute / education
	-0.0481	(-10.4)	Purpose = other
delay10_1-3	-0.291	(-9.2)	Applied to first 3 delayed trains
delay10_4-6	-0.262	(-12.3)	Applied to next 3 delayed trains
delay10_7+	-0.164	(-10.1)	Applied to all delayed trains above 6 Purpose = commute / education
	-0.0861	(-4.2)	Applied to all delayed trains above 6 Purpose = other
delay30_1-3	-0.412	(-14.3)	Applied to first 3 delayed trains
delay30_4+	-0.325	(-10.6)	Applied to all delayed trains above 3
comfort2	-0.305	(-2.2)	Regularity = good
	-0.93	(-6.0)	Regularity = bad
comfort3	-1.73	(-11.2)	
info2	0.0026	(0.0)	Direction = from Paris
	0.287	(4.7)	Direction = towards Paris
info3	0.636	(13.3)	
Cross terms			
timecomf2	-0.0219	(-4.5)	Regularity = good
	-0.0081	(-2.0)	Regularity = bad
timecomf3	-0.0145	(-3.9)	
infody102	0.0609	(3.4)	Direction = from Paris
	0		Direction = towards Paris
tvdelay10	-4.40E-04	(-1.3)	
tvdelay30	-9.10E-04	(-1.5)	
dly30com3	-0.059	(-1.7)	
Scale factors			
Scale3	0.842	(1.5)	
Scale5+6	0.708	(5.4)	
Additional	0.73	(4.8)	
SumTrains	0.912	(0.6)	

Table 4-6 Model parameters and estimated coefficients of the SIXTH model

In a few cases the interval delay coefficients were significantly different from each other. In those cases the coefficient belonging to the lowest interval was the most negative. This means that the extra disutility as a result of one more delayed train out of 20 is decreasing as the number of delayed trains increases. The disutility of the first delayed train out of 20 is high, but the extra disutility of the 8^{th} delayed train is much less.

Some caution needs to be taken when drawing this conclusion. Only people who experienced frequent delays were asked to evaluate service levels with frequent delays. People experiencing only rarely a delay, were offered delay levels between 0 to 4 trains out of 20. The observed difference in the coefficients in the piecewise linear model could also be due to a different valuation of delays by these two groups of respondents. In other words: it might not be at all that a *single* person values the first delayed train out of 20 more negatively than the 8^{th} delayed train. However, when we segmented the respondents according to the frequency of delays that they experience, there was no evidence for any difference in valuation. Furthermore, almost all people were offered frequencies of delay in two frequency-of-delay pieces (in case of the short delays: out of the three pieces 0-3, 4-6 and 7+,). Therefore, the conclusion stated in the previous paragraph is valid.

As was seen in the previous section both delay coefficients were segmented. After making the delay coefficients piece-wise linear, not all the coefficients were still significantly different between the segments. In those cases we have combined the segments again. The resulting model coefficients are listed in Table 4-6 (model SIXTH). The short delay coefficient (delay10) is piecewise-linear over three intervals. The coefficient belonging to the highest interval (7 trains out of 20 delayed and more) is segmented between purpose "commute/education" and purpose "other" as a statistically significant difference still existed here. The long delay coefficient (delay30) is piecewise-linear over two intervals (1-3 trains out of 20 delayed; and 4 and more trains out of 20 delayed).

4.7 Investigation of Distributed Parameters

As a result of the heterogeneous nature of the population of travellers, there could be significant variation in parameter valuations within the population, even after taking account of observed differences through segmentation. Some travellers belonging to the same segment might still value reliability higher or lower than others. The standard logit model is not able to capture this variation, as it assumes that all travellers have the same (constant) utility value for each parameter. One way to test for variation in the parameters is by estimating so-called Error-Components logit models (EC-logit). In such a model additional parameters may be specified which represent the variation in the parameter value: for instance instead of estimating one single coefficient for frequency of short delay (standard logit) one could estimate a coefficient for frequency of short delay *and* an additional coefficient for the standard deviation of frequency of short delay (EC-logit). We have investigated possible random taste variation of the respondents using the EC-logit approach, where we estimate coefficients for both the mean and standard deviation of the distribution of the coefficient values. The introduction of distributional parameters has the potential to increase model explanation and provides important information on the range of coefficient valuations observed in the sample.

Model	SIXTH-B		SIXTH-C		
Observations		20481		20481	
Final log (L)		-13287.7		-13286.5	
D.O.F.		21		27	
Rho²(0)		0.064		0.064	
Rho²(c)		0.062		0.062	
Base terms	***Coeff***	***T-ratio***	***Coeff***	***T-ratio***	***Applied only when***
time	-0.0535	(-20.5)	-0.0535	(-20.5)	Purpose = commute / education
	-0.0391	(-10.1)	-0.0391	(-10.1)	Purpose = other
delay10_1-3	-0.239	(-9.1)	-0.240	(-9.1)	Applied to first 3 delayed trains
delay10_4-6	-0.224	(-12.7)	-0.224	(-12.7)	Applied to next 3 delayed trains
delay10_7+	-0.136	(-10.3)	-0.136	(-10.3)	Applied to all delayed trains above 6 Purpose = commute / education
	-0.0759	(-4.3)	-0.0764	(-4.3)	Applied to all delayed trains above 6 Purpose = other
delay30_1-3	-0.345	(-14.7)	-0.345	(-14.7)	Applied to first 3 delayed trains
delay30_4+	-0.278	(-10.9)	-0.278	(-10.8)	Applied to all delayed trains above 3
comfort2	-0.276	(-2.8)	-0.277	(-2.8)	Regularity = good
	-0.715	(-7.2)	-0.714	(-7.2)	Regularity = bad
comfort3	-1.3	(-14.8)	-1.3	(-14.8)	
info2	0.0132	(0.2)	0.0131	(0.2)	Direction = from Paris
	0.276	(4.8)	0.276	(4.8)	Direction = towards Paris
info3	0.587	(13.2)	0.587	(13.2)	
Cross terms					
timecomf2	-0.014	(-4.6)	-0.014	(-4.6)	Regularity = good
	-0.0044	(-1.6)	-0.0044	(-1.6)	Regularity = bad
timecomf3	-0.0094	(-3.8)	-0.0094	(-3.8)	
infody102	0.0532	(3.1)	0.0531	(3.1)	Direction = from Paris
	0		0		Direction = towards Paris
tvdelay10	-1.60E-04	(-0.6)	-1.60E-04	(-0.6)	
tvdelay30	-2.60E-04	(-0.5)	-2.50E-04	(-0.5)	
dly30com3	-0.0256	(-1.1)	-0.0256	(-1.1)	
Error components / Distributed parameters					
delay10_1-3			-0.0014	(-0.1)	
delay10_4-6			0.1	(0.8)	
delay10_7+			2.7E-04	(0.0)	Purpose = commute / education
delay10_7+			-0.0075	(-0.5)	Purpose = other
delay30_1-3			0.0032	(0.2)	
delay30_4+			-0.0158	(-0.9)	

Table 4-7 Model parameters and estimated coefficients of the SIXTH-B (same as SIXTH, without tree structure) and SIXTH-C (including error components) model

The first difficulty with implementing an EC-logit model with distributed parameters is to replicate the tree structure using error components only (ALOGIT[2] can't estimate models with both tree structures *and* error components simultaneously). Given that all scale coefficients are not very different from one, an error component model was estimated dropping the scaling terms. To compare this model with a model without error components, we also estimated a model that was equal to model SIXTH without tree structure. This model is called SIXTH-B and can be found in Table 4-7.

Tests were undertaken on this model to examine whether improvements could be obtained by specifying the base delay terms within the model as distributed parameters to allow random taste variation. These error components were drawn from a normal distribution using 1000 draws. The error components on these delay terms were not found to have significant coefficient estimates (see Table 4-7, model SIXTH-C), indicating that they added no additional explanatory value to the model specification (it is worth noting that when estimating an error component model, the error component term coefficients have to be considered in terms of their absolute values, the signs are irrelevant). From this it can be concluded that remaining variation in the model is mainly random.

4.8 Correction for Correlation of Responses

An important advantage of an SP experiment is that several responses can be collected from each individual. This reduces substantially the cost of data collection and allows for more advanced experimental designs. However, the collection of multiple responses means that each respondent's basic preferences apply to the series of responses that he or she has given: those responses are therefore interdependent. 'Naïve' analysis methods that assume the independence of observations are therefore in principle invalid.

While a number of methods can be used to correct for the interdependence of SP observations, experience has shown that a good practical method is to use the 'jack-knife' procedure [3, 4] (for a discussion on the application of this technique to SP data, see [5]). This is a standard statistical method for testing and correcting model mis-specifications. RAND Europe has pioneered its use in connection with SP data and has found it to be effective and reliable in this context. In general, the application of the jack-knife procedure to SP data has confirmed that the coefficient estimates themselves are not greatly affected by the specification error of assuming independent observations. However, the significance of the coefficient estimates is often substantially overstated by the naïve estimation. Thus when there is an important issue about the significance of a specific variable it is necessary to test that variable in a jack-knife procedure rather than in a naïve estimation. Generally it is found that when variables are significant at very high levels in a naïve estimation, they remain significant in the jack-knife estimation; but when the significance of a variable in the naïve estimation is marginal, a jack-knife estimation may show that it is not truly significant.

This is indeed what happened when a jack-knife estimation was performed starting from model SIXTH (see Table 4-6). The significance of three terms that were already below the

[2] ALOGIT is the program that is used for doing all model estimations.

threshold (tvdelay10, tvdelay30, dl30com3) decreased even further. The significance of the timecomf2 term (in case the regularity = bad) was just above the threshold. After the jack-knifing it dropped to just below the threshold. We decided to re-estimate the model dropping the first three coefficients, but leaving the last one in.

Also, the scales Scale3 and SumTrains were insignificant (note that they were insignificant in model Sixth as well). This means that there is no longer any evidence that the responses to Game 3 contained a higher level of unexplained variation than the other games (except for games 5 and 6, to which a significant scale factor still applies). Furthermore, it indicates that there is no evidence anymore that the responses of people who said to perceive short or long delays in excess of 20 out of 20 trains contained any more unexplained variation than the responses of other people. The model was re-estimated with these scale factors removed.

Table 4-8 shows the final model with all final T-ratios after the jack-knife. All terms are significant on a 95% confidence level, except the timecomf2 term (for cases where the regularity = bad), which is only significant on a 92.5% confidence level. This model will be used in the application that is part of the next phase of the project.

Model	FINAL		
Observations	20481		
Final log (L)	-13270.4		
D.O.F.	19		
Rho²(0)	0.065		
Rho²(c)	0.064		
Base terms	***Coeff***	***T-ratio***	***Applied only when***
time	-0.0637	(-16.8)	Purpose = commute / education
	-0.0476	(-8.6)	Purpose = other
delay10_1-3	-0.293	(-9.4)	Applied to first 3 delayed trains
delay10_4-6	-0.265	(-11.1)	Applied to next 3 delayed trains
delay10_7+	-0.164	(-14.9)	Applied to all delayed trains above 6 Purpose = commute / education
	-0.0905	(-4.6)	Applied to all delayed trains above 6 Purpose = other
delay30_1-3	-0.424	(-14.6)	Applied to first 3 delayed trains
delay30_4+	-0.342	(-16.6)	Applied to all delayed trains above 3
comfort2	-0.311	(-2.1)	Regularity = good
	-0.892	(-5.8)	Regularity = bad
comfort3	-1.73	(-12.8)	
info2	0		Direction = from Paris
	0.283	(5.9)	Direction = towards Paris
info3	0.628	(11.4)	
Cross terms			
timecomf2	-0.0201	(-3.8)	Regularity = good
	-0.0068	(-1.8)	Regularity = bad
timecomf3	-0.0125	(-4.5)	
infody102	0.0605	(3.3)	Direction = from Paris
	0		Direction = towards Paris
Scale factors			
Scale5+6	0.736	(4.1)	
Additional	0.731	(5.1)	

Table 4-8 Model parameters and estimated coefficients of the final model

REFERENCES

Reference List

[1] Eric Kroes, Hugues Duchâteau, *Questionnaires recrutement et SP*, Note de Travail, 27 April 2004

[2] Laurence Debrincat, *Note 040604*, distributed by e-mail on 4 June 2004

[3] Bissell, A.F. and R.A. Ferguson *The Jackknife: Toy, Tool or Two-Edged Weapon?.* 1975, Statistician. V. 24. pp79-100

[4] Miller, R.G. *The Jackknife: A Review,* 1974, Biometrica. V. 61. pp 1-14.

[5] C. Cirillo, A. J. Daly & K. Lindveld *Eliminating Bias due to the Repeated Measurements Problem in SP Data* in *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, 1998, M. Labbé et al. (eds.), Springer.

APPENDICES

Appendix A: Recruitment Questions

The following tables give an overview of the responses of the 1273 respondents to the recruitment questions (this is before the exclusion of the respondents that gave a non-intuitive answer to the dominant question).

Q: Combien de temps en minutes va durer votre trajet jusqu'à <ARRIVAL> (for departing passengers; analogue question for arriving passengers)

	Frequency	Percent
10 - 12 min	173	13.6
13 - 17 min	191	15.0
18 - 22 min	164	12.9
23 - 27 min	127	10.0
28 - 32 min	149	11.7
33 - 37 min	85	6.7
38 - 42 min	103	8.1
43 - 47 min	109	8.6
48 - 52 min	29	2.3
53 - 57 min	10	.8
58 - 64 min	69	5.4
65 - 74 min	19	1.5
75 - 84 min	18	1.4
85 - 94 min	24	1.9
95 - 104 min	1	.1
105 - 114 min	2	.2
Total	**1273**	**100.0**

Remark: the respondent could specify his travel time up to the minute. These reported travel times have been clustered into 5 and 10 minutes interval to create this table.

Q: Aujourd'hui, quel est le motif principal de votre déplacement ?

	Frequency	Percent
Domicile -> travail	446	35.0
Travail -> domicile	258	20.3
Domicile -> études	119	9.3
Études -> domicile	82	6.4
Autres motifs	368	28.9
Total	**1273**	**100.0**

Q: Selon vous, sur le trajet que vous êtes en train de faire, combien de trains sur 20 arrivent avec 5 à 15 minutes de retard ?

	Frequency	Percent
0	133	10.4
1	93	7.3
2	206	16.2
3	170	13.4
4	127	10.0
5	206	16.2
6	42	3.3
7	27	2.1
8	23	1.8
9	9	.7
10	147	11.5
11	2	.2
12	15	1.2
13	7	.5
14	3	.2
15	38	3.0
16	1	.1
17	2	.2
18	11	.9
19	4	.3
20	7	.5
Total	**1273**	**100.0**

Q: Selon vous, sur le trajet que vous êtes en train de faire, combien de trains sur 20 arrivent avec 15 minutes ou plus de retard ?

	Frequency	Percent
0	443	34.8
1	320	25.1
2	200	15.7
3	80	6.3
4	42	3.3
5	97	7.6
6	18	1.4
7	11	.9
8	20	1.6
9	1	.1
10	32	2.5
11	2	.2
12	2	.2
13		
14		
15	4	.3
16		
17		
18		
19		
20	1	.1
Total	**1273**	**100.0**

Q: Sur ce trajet que vous êtes en train d'effectuer, diriez-vous que vous voyagez ...

	Frequency	Percent
toujours assis	755	59.3
le plus souvent assis	387	30.4
le plus souvent debout	97	7.6
le plus souvent debout et serré	19	1.5
toujours debout et serré	15	1.2
Total	**1273**	**100.0**

Q: Vous êtes:

	Frequency	Percent
Homme	614	48.2
Femme	659	51.8
Total	**1273**	**100.0**

As a check we have looked at whether respondents travelling with lines that are marked 'Mauvaise Regularité" are really perceiving a higher number of trains delayed. This is indeed the case, as can be seen from the following table.

	Trains out of 20 with a delay between 5 and 15 minutes		Trains out of 20 with a delay of 15 minutes or more	
	Mean	Std. Deviation	Mean	Std. Deviation
Bonne Régularité	3.9	3.8	1.4	2.2
Mauvaise Régularité	5.5	4.2	2.3	2.6
Total	**4.8**	**4.1**	**1.9**	**2.5**

As another check we have looked into the comfort levels. The observed number of people who could always sit in the train was higher than expected. This can be explained by the high number of people that were recruited outside the peak hours. The following table gives the distribution over the comfort levels of the respondents grouped into clusters that were recruited at about the same time-of-day. This table shows that the fraction of people in each comfort class depends on the time-of-day that they travelled.

	ACTUALCOMFORT					Total
	toujours assis	le plus souvent assis	le plus souvent debout	le plus souvent debout et serré	toujours debout et serré	
0:00 - 7:29	83 *82.2%*	10 *9.9%*	6 *5.9%*	0 *.0%*	2 *2.0%*	101 *100.0%*
7:30 - 9:29	140 *61.4%*	64 *28.1%*	16 *7.0%*	5 *2.2%*	3 *1.3%*	228 *100.0%*
9:30 - 14:59	276 *56.2%*	162 *33.0%*	41 *8.4%*	7 *1.4%*	5 *1.0%*	491 *100.0%*
15:00 - 16:59	88 *49.4%*	68 *38.2%*	14 *7.9%*	4 *2.2%*	4 *2.2%*	178 *100.0%*
17:00 - 18:59	132 *62.0%*	61 *28.6%*	16 *7.5%*	3 *1.4%*	1 *.5%*	213 *100.0%*
19:00 - 23:59	36 *59.0%*	22 *36.1%*	3 *4.9%*	0 *.0%*	0 *.0%*	61 *100.0%*
Not recorded	0 *.0%*	0 *.0%*	1 *100.0%*	0 *.0%*	0 *.0%*	1 *100.0%*
Total	755 *59.3%*	387 *30.4%*	97 *7.6%*	19 *1.5%*	15 *1.2%*	1273 *100.0%*

Interviews were done with people travelling at the following lines:

	Frequency	Percent
Paris Est - Bondy-Aulnay	31	2.4
Paris Est - Château Thierry	8	0.6
Paris Est - Coulommiers - La Ferté	11	0.9
Paris Est - Esbly-Crécy	36	2.8
Paris Est - Longueville	17	1.3
Paris Montparnasse - Dreux	22	1.7
Paris Montparnasse - Plaisir-Mantes	25	2.0
Paris Montparnasse - Rambouillet	25	2.0
Paris Montparnasse - Sèvres	36	2.8
Paris Nord - Crépy	27	2.1
Paris Nord - Persan (via Montsoult)	55	4.3
Paris Nord - Persan (via Valmondois)	31	2.4
Paris Nord - Pontoise	20	1.6
Paris St Lazare - Cormeilles en Parisis	27	2.1
Paris St Lazare - La Verrière, La Défense, St Quentin	9	0.7
Paris St Lazare - Maisons Laffitte, Nanterre U, Cergy	39	3.1
Paris St Lazare - Mantes via Conflans, Gisors	26	2.0
Paris St Lazare - Mantes via Poissy	25	2.0
Paris Sud Est - Montereau	17	1.3
RER A	60	4.7
RER B	73	5.7
RER C - Brétigny	127	10.0
RER C - Dourdan	19	1.5
RER C - Etampes	49	3.8
RER C - Massy	62	4.9
RER C - St Quentin	110	8.6
RER C - Versailles	8	0.6
RER C - Versailles Chantiers	8	0.6
RER D - Nord	55	4.3
RER D - SE (Combs)	35	2.7
RER D (Corbeil)	57	4.5
RER E - Chelles	68	5.3
RER E - Villiers	55	4.3
Total	**1273**	**100.0**

Appendix B: Background Questions

The following tables give an overview of the responses of the 1273 respondents to the background questions (this is before the exclusion of the respondents that gave a non-intuitive answer to the dominant question).

Q: Toujours en pensant au trajet dont nous venons de parler, que diriez-vous du niveau d'information :

	Frequency	Percent
On vous annonce qu'il y a une perturbation	501	39.4
On vous annonce qu'il y a une perturbation et quelle en est la cause	393	30.9
On vous annonce qu'il y a une perturbation, quelle en est la cause et l'ampleur du retard prévu	378	29.7
Total	**1272**	**99.9**
No answer	1	.1
Total	**1273**	**100.0**

Q: Lorsque vous effectuez le trajet pour motif <PURPOSE>, à l'arrivée, avez-vous:

	Frequency	Percent
Un horaire impératif	584	45.9
Un horaire modulable à 15 minutes près	431	33.9
Un horaire modulable à 30 minutes près	109	8.6
Un horaire modulable à 1 heure près	42	3.3
Aucun impératif horaire	107	8.4
Total	**1273**	**100.0**

Q: Toujours en pensant au trajet dont nous venons de parler, aviez-vous un horaire de correspondance à respecter :

	Frequency	Percent
Oui, j'ai fait une correspondance train / RER après	499	39.2
Oui, j'ai fait une correspondance bus après	202	15.9
Non, je n'avais pas d'horaire de correspondance à respecter	572	44.9
Total	**1273**	**100.0**

Q: Quel titre de transport utilisez-vous pour faire ce trajet de de <DEPARTURE> à <ARRIVAL> :

	Frequency	Percent
Carte Orange mensuelle ou hebdomadaire	590	46.3
Ticket à l'unité ou en carnet	148	11.6
Imagine'R / Navigo	312	24.5
Intégrale / Navigo	191	15.0
Autres titres	32	2.5
Total	**1273**	**100.0**

Q: Vous effectuez ce même trajet :

	Frequency	Percent
Tous les jours	907	71.2
3 fois par semaine	191	15.0
1 fois par semaine	88	6.9
moins souvent	87	6.8
Total	**1273**	**100.0**

Q: Vous arrive-t-il de faire ce trajet, ne serait-ce que de temps en temps en voiture ?

	Frequency	Percent
Oui	539	42.3
Non	734	57.7
Total	**1273**	**100.0**

Q: Quel est votre âge ?

	Frequency	Percent
15 à 19 ans	167	13.1
20 à 29 ans	501	39.4
30 à 39 ans	254	20.0
40 à 49 ans	188	14.8
50 à 59 ans	134	10.5
60 ans et plus	29	2.3
Total	**1273**	**100.0**

Appendix C: Background to the model development

The model is based on the assumption that the respondent chooses the alternative with the highest utility. An error term is included in the utility function to reflect the unobservable factors in the individual's utility. The estimation can therefore be conducted within the framework of random utility theory, i.e. accounting for the fact that the analyst has only imperfect insight into the utility functions of the respondent.

The most popular and widely available estimation procedure is logit analysis. The logit model predicts choice probabilities as

$$P_1 = \exp U_1 / (\exp U_1 + \exp U_2 + \exp U_3),$$

where the U's represent the utility functions of the alternatives and exp is the standard exponential function. The estimation procedure produces estimates of the model coefficients, such that the choices made by the respondents are best represented. Parameters are added and removed until a satisfactory model for the utilities is generated. The standard statistical criterion of Maximum Likelihood is used. Both the values of the coefficients (in utility terms) and the significance of the coefficients are output.[3]

The estimation of the discrete choice models was undertaken using ALOGIT. For each model, two sets of values are presented: (i) model summary statistics, and (ii) model coefficients and their associated approximate t-ratios.[4]. The model summary statistics are defined as follows:

Observations
The number of observations included in the model estimation.

Final log (L)
This indicates the value of the log-likelihood at convergence. The log-likelihood is defined as the sum of the log of the probabilities of the chosen alternatives, and is the function that is maximised in model estimation. The value of log-likelihood for a single model has no obvious meaning. However comparing the log-likelihood of two

[3] For further information about logit models, see Ben-Akiva, M. and Lerman S. R. (1985) Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge, Massachusetts.

[4] This ratio is an asymptotic approximation to the standard statistical Student's T-ratio.

models with different specifications allows the statistical significance of new model coefficients to be assessed properly.

D.O.F.

Degrees of freedom, i.e. the number of coefficients estimated in this model. Note that if a coefficient is constrained to a fixed value (indicated by(*)) then it is not a degree of freedom.

Rho²(0)

The rho-squared measure compares the log-likelihood (LL(final)) to the log-likelihood of a model with all coefficients restricted to zero (LL(0)):

$$Rho^2(0) = 1 - LL(final)/LL(0)$$

A higher value indicates a better fitting model.

Rho²(c)

If we compare the log-likelihood (LL(final)) value obtained with the log-likelihood of a model with only constants (LL(c)) we get:

$$Rho^2(c) = 1 - LL(final)/LL(c)$$

Again a higher value indicates a better fitting model.

The coefficient values are then presented. If a coefficient is positive then it has a positive impact of utility and so reflects a higher probability of choosing the alternatives to which it is applied, for example, better comfort. Conversely if a coefficient is negative then it has a negative impact on utility and so reflects a lower probability of choosing the alternative to which it is applied, for example, increased travel time.

Some coefficients are multiplied by continuous variables and therefore reflect the disutility per unit of the variable, e.g. time, which reflect the relative disutility per minute of travel time. Other coefficients are applied to categorical variables; these therefore reflect the total utility increase or decrease for that variable, relative to a base situation, e.g. the increase in utility as a result of having a certain comfort level. In some cases, significant coefficients could not be identified for each discrete level for a variable and therefore valuations for some levels have been aggregated.

The constants in each model reflect preferences for the alternatives to which they are applied. The constants on the models are additive and more than one constant can be applied for each individual. A positive value for a constant indicates that the respondent is more likely to choose that alternative, and a negative value for a constant indicates that the respondent is less likely to choose that alternative.