DESIGN OF DATA COLLECTION PROGRAMS

Outline

1. Overall Design
2. Direct vs. Indirect Measurement
3. Data Variability
4. Sample Size Equations
5. Program Design Process
6. Survey Design

Data Collection Program Elements

A. **Baseline:**
- Develop route profiles
- Define base conditions
- Develop conversion factors

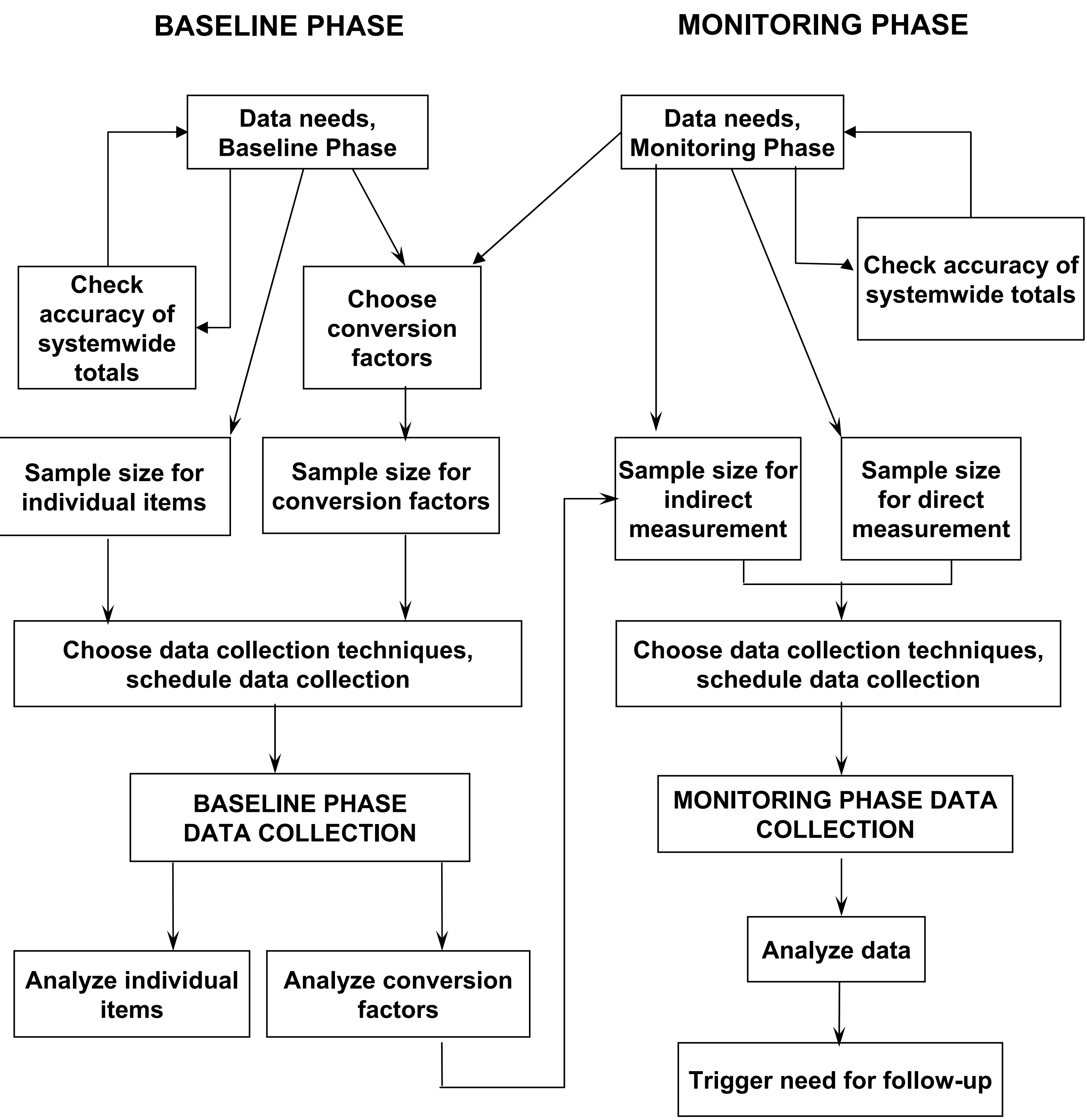
B. **Monitoring:**
- Detect changes based on selective data collection
- Use conversion factors to estimate other data

C. **Follow Up:**
- Develop new route profiles
- Selective additional data
- Special studies

Conversion Factors

Auxiliary Data Item	Inferred Data Item
Load or Revenue	Boardings
Boardings, Load or Revenue	Passenger Miles
Point Load	True Maximum Load
Revenue	Peak Point Load

Designing a Data Collection Program



Default Values for Coefficient of Variation of Key Data Items

Data Item	Time Period	Route Classification	Default Value
Maximum Load	Peak	< 35 pass./trip	0.5
		≥ 35 pass./trip	0.35
	Off- Peak	< 35 pass./trip	0.6
		35-55 pass./trip	0.45
		> 55 pass./trip	0.35
	Evening	All	0.75
	Owl*	All	1
	Sat, 7 AM-6 PM	All	0.6
	Sat, 6 PM-1 AM	All	0.75
	Sun, 7 AM-1 AM	All	0.75
Boardings, Passenger Miles	Peak	< 35 pass./trip	0.42
		≥ 35 pass./trip	0.35
	Off- Peak	< 35 pass./trip	0.45
		35-55 pass./trip	0.4
		> 55 pass./trip	0.35
	Evening	All	0.73
	Owl*	All	0.8
	Sat, 7 AM-6 PM	All	0.45
	Sat, 6 PM-1 AM	All	0.73
	Sun, 7 AM-1 AM	All	0.73
Running Time	All	short (≤ 20 min.)	0.16
		long (> 20 min.)	0.1

*Owl default values are the same for weekdays and weekends

Inherent Variability Example

A paired sample: boardings, pass-mi by trip

	Brdgs	Pass-mi
Sample #1	10	30
Sample #2	30	150
Total	40	180
Mean	20	90

avg trip length (ATL) = 90/20 = 4.5 mi

Direct estimation

Best pass.-mi. estimate is the mean.

Deviations from "best estimate"

Sample #1 30-90 = -60
Sample #2 150-90 = +60

Ratio estimation

Best pass.-mi. estimate is brdgs * ATL

Deviations from "best estimate"

Sample #1 30 - (10*4.5) = -15
Sample #2 150 - (30*4.5) = +15

Coefficient of Variation: Measure of Inherent Variability

cv **= standard deviation / mean**

$n = (2 * cv)^2 / (\text{target precision})^2$

So doubling *cv* means quadrupling the necessary sample size

- **Direct estimation approach:**

  cv **= 85/60 = 1.3 *(very large!)***

  n **= 676 (assumes target precision = 10%)**

- **Ratio estimation approach:**

  cv **= 21/60 = 0.35 *(nice and small!)***

  n **= 50**

Pass.-Mi. Sampling Techniques: Santa Cruz Case Study

	unit cv	n
FTA Circular 2710.1A (3 trips every other day)	n.a.	549
Direct estimation	0.95	522
Ratio to boardings	0.72	296
Ratio to boardings, 4 strata	0.43	117
Ratio to (boardings * rt length)	0.44	112

Sampling Strategies

Simple random sampling

Every trip has equal likelihood of selection

Systematic sampling

Sample every 6th day – like random, but avoid cycles
Smoothes data collection load
Example: ***FTA Circular***

Cluster sampling

Identify natural clusters in advance, select them at random
With passenger surveys, bus trip = cluster of passengers
Example: ***on-board survey***
Example: ***sample round trips, or clusters of 4 trips***

Ratio estimation

Take advantage of complete or less expensive data sources
Example: ***convert farebox boardings to pass.-mi***
Example: ***convert load at checkpoint to load elsewhere***

Stratified sampling

Separate sample for each stratum
Example: ***long vs. short routes for average trip length***

Two-Stage Variance and Two-Stage Sampling

Average running time on Route X during period Y is affected by:

- **day-to-day variation**
- **scheduled trip - to - scheduled trip variation (e.g., the 7:45 takes longer than the 7:55)**
- **random variation**

If it is convenient to sample in concentrated periods (e.g., every trip in one day), it's a two-stage sample

- **day-to-day variation reduced by sampling over several days**
- **trip - to - trip variation and random variation reduced by sampling many trips and many days**

Caution: a sample conducted on only one day or a small number of days may still have a lot of variation, and consequently the estimate will have a wide precision.

Two-stage sampling can also be done for boardings, pass.-miles on individual lines (e.g., light rail line), or even for entire system (e.g, FTA circular 2710.1A – but it's overkill)

Sample Size Equations

Simple Random Sample:

$$n = \frac{3.24\, v^2}{d^2} \quad \text{or} \quad d = \frac{1.8\, v}{\sqrt{n}}$$

Where n = sample size (number of trips)
d = tolerance (e.g. d = .05 means ± 5% tolerance)
v = coefficient of variation
90% confidence level assumed

Required Sample Size for Estimating Averages

	d = tolerance									
v	0.5	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	13	4	2	1	1	1	1	1	1	1
0.2	52	13	6	4	3	2	2	1	1	1
0.3	117	30	13	8	5	4	3	2	2	2
0.4	208	52	24	13	9	6	5	4	3	3
0.5	324	82	36	21	13	10	7	6	5	4
0.6	467	117	52	30	19	13	10	8	6	5
0.7	636	159	71	40	26	18	13	10	8	7
0.8	830	208	93	52	34	24	17	13	11	9
0.9	1050	263	117	66	42	30	22	17	13	11
1	1296	325	144	82	52	37	27	21	17	13
1.25	2025	507	225	127	82	57	42	32	25	21
1.5	2917	730	324	183	117	82	60	46	37	30

Notes: assuming 90% confidence level
v = coefficient of variation

Conversion Factor Equations

Compute conversion factor and its coefficient of variation:

$$R = \frac{\bar{y}}{\bar{x}}$$

Where R = conversion factor

$\bar{y}$ = **average of inferred data item (e.g. boardings) in paired sample**

$\bar{x}$ = **average of auxiliary data item (e.g. load) in paired sample**

$$V_R^{\ 2} = \frac{1}{n-1.7}(V_x^{\ 2} + V_y^{\ 2} - 2V_x V_y r_{xy})$$

Where V_R = **C.O.V. of conversion factor**

V_x = **C.O.V. of auxiliary item (e.g. load)**

V_y = **C.O.V. of inferred item (e.g. boardings)**

r_{xy} = **correlation coefficient between auxiliary and inferred items**

Also n = **number of paired observations in sample**

$$S_{xy} = COV(x, y) = \frac{(\sum x_i y_i) - n\overline{xy}}{n-1}$$

$$and \quad r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Determine Sample Size in Monitoring Phase

$$n_2 = \frac{V_x^2(1+V_R^2)}{0.31d_m^2 - V_R^2}$$

Where n_2 **= sample size of auxiliary item in monitoring phase**

d_m **= desired tolerance of the inferred data item**

b. Desired Tolerance of Inferred Item = ±10%

V_R^2 / V_x	.0001	.0005	.001	.0015	.002	.00225	.0025	.00275
0.10	4	4	5	7	10	12	17	29
0.20	14	16	20	26	37	48	67	115
0.30	31	35	43	57	82	107	151	258
0.40	54	62	77	101	146	189	268	459
0.50	84	97	120	157	228	295	418	717
0.60	121	139	172	226	328	425	602	1032
0.70	16	189	234	307	447	578	819	1404
0.80	214	247	306	401	583	755	1070	1834

Notes: assuming 90% confidence level

V_x **= coefficient of variation of auxiliary item**

V_R^2 **= square of coefficient of variation of conversion factor**

Step-by-Step Data Collection Program Design Procedure

1. <u>Determine data needs and acceptable tolerances</u> based on uses of data

2. Select <u>statistical inputs</u> (i.e. coefficient of variation) based on preliminary data analysis and/or default values.

3. Select <u>data collection techniques</u> based on data needs and efficiency of each technique for property.

e.g.	Baseline:	ridechecks + supplementary point checks
	Monitoring:	pointchecks
	Update:	ride checks

4. Determine <u>constraining sample sizes</u> for each technique by route and time period by applying formula.

5. Determine detailed <u>checker requirements</u> for each route and time period.

6. <u>Estimate ratios</u> (e.g. average fare, trip length, peak load/total passengers) using baseline data for possible use in monitoring.

7. <u>Revise monitoring plan</u> (techniques and sample sizes) based on data analysis.

Sample Size for Proportions

Using absolute equivalent tolerance (AET),

$$n = .96/AET^2$$

Recall conversion from tolerance around an expected proportion p to AET:

$$AET = 0.5\ tol\ /\ sqrt[p*(1-p)],$$

where *p* = expected proportion

Rule of thumb: LARGE sample size needed to estimate a proportion accurately!

n	600	267	150	96
AET	4%	6%	8%	10%
p = 50%	4%	6%	8%	10%
p = 60%	3.9%	5.9%	7.8%	9.8%
p = 70%	3.7%	5.5%	7.3%	9.2%
p = 80%	3.2%	4.8%	6.4%	8.0%
p = 90%	2.4%	3.6%	4.8%	6.0%
p = 95%	1.7%	2.6%	3.5%	4.4%

Sample Size for Passenger Surveys

- **Determine needed sample size for proportion**

  e.g., proportion of passengers who are pleased, who own a car, etc.

- **Multiply SS if proportions are desired for various strata**

  e.g., proportion of passengers car-owning passengers who are pleased

- **Multiply by "clustering effect"**

  e.g., in on-board survey, 4 responses from same bus may be equivalent to 1 response from a randomly selected rider; clustering effect depends on question

  if so, expand SS by 4

- **For origin-destination matrix,**

  SS = 20 * number of cells (rule of thumb)

  level of detail determined number of cells

- **Expand by 1/(response rate)**

- **Be prepared to revise your expectations when you see how large the needed sample is!**

Response Rate

Along with getting correct answers, your primary concern should be getting a high response rate

- **Cost: lower response rate means more surveying to get the needed number of responses**
- **Non-response bias: non-responders may be different from responders, *and you'll never know!***

Some non-response bias is predictable and insidious:

- **standees are less likely to respond, making close-in origins underrepresented**
- **low literacy, teens, & non-native population respond less**
- **predicable biases can be modeled and corrected by numerical procedures**

Ways to improve response rate:

- **shorten the questionnaire**
- **quick oral survey: "What station are you going to?"**
- **get info from counts whenever possible (e.g., fare type)**
- **distribution method, surveyor training, supervision**
- **believe and experiment!**

The Survey Design Process

1. Define survey objectives
2. Define the population to be surveyed
3. Determine data requirements
4. Specify precision required
5. Select survey instrument
6. Define sampling unit
7. Select sampling procedure and sample size
8. Pretest the survey
9. Develop the survey management process
10. Determine analysis methods
11. Develop data storage and management system

Questionnaire Design

- Length
- Layout
- Instructions
- Questions

Questions need to be:

- Understandable
- Answerable
- Well-motivated
- Useful
- Exhaustive

Questions must not be:

- Double-barreled
- Use double negatives
- Use technical jargon
- Long-winded
- Biased
- Redundant
- Self-evident
- Overly intrusive
- Make the respondent uncomfortable

Pre-testing the Survey

- One of the single most important steps in the entire process
- Far better to make mistakes on a small pretest than on the full survey

Survey Management

Especially for large surveys, the survey management process is critical to the successful execution of the survey

- Quality control
- Response rate
- Cost control

Includes:

- Recruitment and training of interviewers
- Supervision of survey staff
- Procedures for data capture and cleaning
- Communications with the public

Equally important is the data storage and management issues once the survey is over and the data have been obtained

- Documentation of procedures, files, etc.
- Database management systems -- data accessibility
- Tying into other databases
- Keeping the database "alive" and useful

Onboard Surveys

Advantages:

- Cheap, easy to administer
- Efficient means of obtaining information on current riders
- Before-and-after study of service changes

Disadvantages:

- Limited information obtained
- No information concerning non-users

Major Design Considerations:

- Distribution/collection of the survey forms
- Keeping questions asked to a minimum

Mailback Surveys

Advantages:

- Cheap
- Fairly extensive information can be collected
- Household-based sampling frame well defined
- Can get at non-transit users, etc.

Disadvantages:

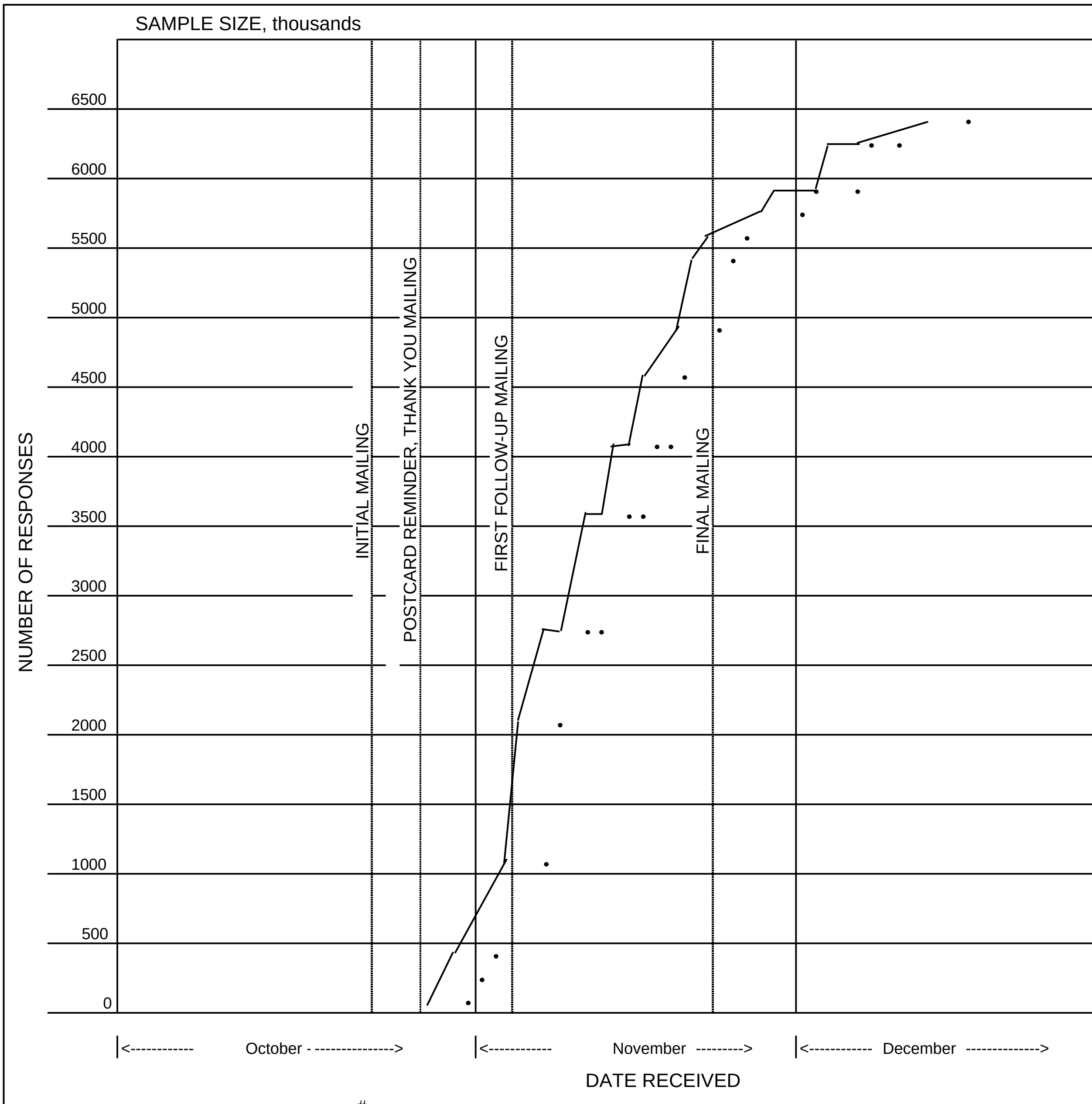
- Low response rates unless very well designed and executed

Major Design Considerations:

- Must do everything possible to achieve a high response rate!
- Can be combined with telephone methods

1993 Scarborough Transit Services Improvement Study

Cumulative Daily Responses



Telephone Surveys

Advantages:

- Efficient use of interviewers
- Good supervisory control
- Good response rates
- Reasonable amounts of information obtained

Disadvantages:

- Respondents must have a telephone
- Call screening
- Bias with respect to frequent travelers
- Limitations on question complexity
- Potential for "respondent bias"

Major Design Considerations:

- Initial contact
  - "cold calling"
  - advance letter
- Computer-aided telephone interviewing (CATI)
- Script design; interviewer quality & training

1978 Montreal Telephone Survey Response Rates

Successful Interviews (72.1%)

1st call:	44.2%
2nd call:	19.6%
3rd call:	6.1%
4th call:	2.2%
Total	72.1%

Unsuccessful Interviews (27.9%)

Phone out of service	3.7%
Incomplete response	4.3%
Refused interview	11.6%
No contact	8.3%
Total	27.9%

Types of Surveys

"Revealed Preference"

- Information on actual choices made is gathered
- Most surveys are of this type

"Stated Preference"

- People are placed in hypothetical choice situations and asked what they would do if they were faced with this particular choice
- Stated preference survey methods in transportation have evolved and improved considerably over the past decade. Now generally considered a "proven methodology."

  [J. Polak & P. Jones, "Using Stated-Preference Methods to Examine Traveler Preferences and Responses," in P.R. Stopher and M.E.H. Lee-Gosselin (eds.) *Understanding Travel Behaviour in an Era of Change*, Oxford: Pergamon-Elsevier, 1995.]

Attitudinal surveys

- Rather than ask what people did or might do, these surveys focus on <u>why</u> people do the things they do, and <u>how</u> they feel about the services available to them, etc.

Treatment of Time

Most surveys are cross-sectional; that is, they gather information about travel behavior, etc. at a single point in time.

- ***Cannot observe changes in behavior in response to changes in system conditions, etc. That is, system dynamics cannot be directly observed.***

In order to observe dynamics at work, one needs time-series data; that is, observations of the system (and, ideally, the same people) over time.

- **repeated cross-sections**
- **retrospective surveys**
- **panel surveys**

TTC Panel Survey

- Panel recruited from the study route catchment area
- Each panel member was asked to record all trips made during a 2-week period before, and a 2-week period after, a service change.
- Simple diary format
- Incentives: lottery ticket at recruitment
weekly draws for cash prizes
- Respondents required to mail back weekly travel diaries
- Weekly telephone contacts

Panel Survey Results: Mt. Pleasant Pilot Test

- **72% of eligible people contacted agreed to participate**
- **75% of original panel completed all 4 weeks of the survey**
- **50% increase in peak-period headway + 100% increase in early evening headway resulted in:**
  - **17% decrease in Mt. Pleasant route ridership (1.3 trips/person/week decrease)**
  - **Other transit routes' ridership increased by 0.7 trips/person/week**
  - **Workers shifted routes, not modes, for trips**
  - **Workers shifted modes for non-work trips**
  - **Non-workers (principally seniors) made fewer trips**

Example Page from TTC Panel Survey Trip Diary

Diary for trips beginning at home (Friday, October 2, 1987)

Other Modes	20	40	60	80	100	120
Eglinton	19	39	59	79	99 1	119 1
Davisville	18	38	58	78	98	118
Mt. Pleasant						
Other Modes						
Eglinton	15	35	55	75	95	115
Davisville	14	34	54	74	94	114
Mt. Pleasant						
Other Modes						
Eglinton	11	31	51	71	91	111
Davisville	10	30	50	70	90	110
Mt. Pleasant						
Other Modes						
Eglinton	7	27	47	67	87	107
Davisville	6	26	46	66	86	106
Mt. Pleasant						
Other Modes						
Eglinton	3	23	43	63	83	103
Davisville	2	22	42	62	82	102
Mt. Pleasant						