Assessment of a comprehensive data validation
of a combined mobility-activity-expenditure survey

Florian Aschauer[1]*, Regine Gerike[2], Reinhard Hössinger[1]

[1] Institute for Transport Studies, University of Natural Resources and Life Sciences, Vienna, Austria
[2] Technische Universität Dresden, Chair of Integrated Transport Planning and Traffic Engineering
* corresponding author florian.aschauer@boku.ac.at

Abstract submitted for the *7th symposium arranged by European Association for Research in Transportation (hEART)*, to be held September 2018 in Athens, Greece.

Key words: data validation, item non-response, travel survey methods, time use, data collection, data quality, survey quality

1 Motivation and objectives

Item non-response is a critical issue in diary-based (National-) Travel Surveys (NTS) and Time Use Surveys (TUS) (Gerike et al. 2015, Aschauer et al. 2018). The non-response problem increases if sophisticated analysis techniques shall be applied, most of which rely on additional information to be reported. Examples are (i) analyses in GIS, which require detailed addresses of the origin and destination of each trip, and (ii) mode choice models, which require alternative-specific attributes to be annotated to each reported trip. Brög et al. (1982) distinguishes three main types of non-response errors:

1) errors that could be detected during data processing without additional observed information (e.g. missing addresses of destinations that are visited at least twice in the reporting period);
2) errors that could only be detected by means of follow-up exploration (e.g. missing trips); and
3) errors that could not be detected, because the respondent is unwilling to disclose the information.

While type (i) errors can be corrected without re-contacting of respondents and type (iii) errors cannot be corrected at all, errors of type (ii) could be corrected by immediate data validation procedures, which should be seen as an integral part of conducting a survey (Richardson et al. 1995). There is little doubt about the effectiveness of ex-post validation, but a detailed item-by-item comparison of pre- and post-validation data is still missing, as is detailed information about the expected effort.

The current study contributes to this strand of research by means of two innovations: (i) Using a combined survey of travel behaviour and time use, we analyse and compare the effect of validation for these two types of data as well as for personal data (sociodemographic characteristics, mobility tools) on survey data quality. (ii) We analyse for the first time the effort of validation and provide an indicator for the expected cost-effectiveness of the validation process.

2 Data

The data used for the analysis were obtained from a paper-and-pencil Mobility-Activity-Expenditure Diary (MAED), a novel survey format, which combines three survey disciplines into an integrated diary format: a travel survey, time use survey, and a consumer expenditure survey. The MAED survey yielded a representative sample of 416 Austrian workers; it includes all travel decisions, activities and expenditures for each respondent over a period of one week. During the field work of the survey we took two precautions in view of the validation analysis: (i) we 'froze' all recorded data of a person prior to validation (comprising only the information that could be extracted from the questionnaires); and (ii) the members of the survey team recorded all person hours separately, which were assigned to the ex-post validation.

The validation procedure included an automatically generated, detailed error protocol describing the missing information from the questionnaire, after it has been entered in a database and checked for Type 1) non-response errors. Subsequently, respondents where called back by the survey staff using the protocol in order to retrieve the missing information. A €40 Incentive was sent to them by mail after their questionnaire has been validated over the phone. Three quarters of all survey participants received a validation call.

3 Analysis

The following descriptive analyses are presented in the paper: (i) number of missing items: missing personal items, trips and activities; (ii) trips and activities with missing or wrong information: missing or wrong mode and activity type, other missing or wrong key information, which is required in data analysis; (iii) the characteristics of missed trips and activities are analysed in post-validation data using standard indicators. Significant differences between pre- and post-validation data are tested for.

4 Preliminary results

Travel data:

The (post-) validated dataset contains 1.9 % additional trips and 3.5 % trips with added missing modes (see Table 1). This results in under-reporting of car-driver trips (−3.2 %) and walking trips (−1.4 %). These values are lower than expected. One possible reason is that MAED has in general more reported trips than a conventional travel diary (see Aschauer et al. 2018), mainly because reporting of trips is embedded in reporting of all daily (travel and non-travel) activities; this makes it less likely that trips are forgotten or deliberately omitted. The share of trips added in the validation process increases when only valid trips from the pre-validation dataset are considered (leaving only matchable trip IDs). 160 trips were removed in the validation phase because these were wrongly reported ones. Compared to these 9,524 trips with matchable trip IDs, the validated dataset contains 3.5 % additional trips but less additional trips with missing mode (2.9 %).

Table 2 shows that 27 % of trips had at least one missing or wrong key information, which was required for data analysis. The major part refers to missing or wrong addresses. Moreover, personal characteristics were also often missing, which were required for the annotation of alternative-specific attributes for the mode choice model (license, availability of public transport season ticket etc.). A detailed analysis of missing person-related information as well as of the characteristics of missed trips will be included in the final paper.

Table 1: Trips by mode in pre- and post-validation data

	No. of trips			share [%]			difference [%]	
Travel mode	post-val.	pre-val.	pre-val.*	post-val.	pre-val.	pre-val.*	pre-val.	pre-val.*
missing	25	350	285	0.3	3.5	2.9	*3.3*	*2.6*
walk	1,295	1,158	1,141	13.1	11.7	11.6	*-1.4*	*-1.6*
bike	415	416	409	4.2	4.2	4.1	*0*	*-0.1*
moto	57	57	57	0.6	0.6	0.6	*0*	*0*
car-driver	6,266	5,953	5,908	63.5	60.3	59.8	*-3.2*	*-3.6*
car-passenger	854	828	823	8.6	8.4	8.3	*-0.3*	*-0.3*
public	961	922	901	9.7	9.3	9.1	*-0.4*	*-0.6*
Total	**9,873**	**9,684**	**9,524**	**100**	**98.1**	**96.5**		

* pre-validated data with trips of equal trip ID

Table 2: Trips with missing or wrong key information in pre-validation data

Indicator	no. of trips	percent
Total trips	**9,684**	**100.0**
Invalid trips (trip ID not matchable to post-validation)	160	1.7
Trips with missing or different mode	315	3.3
Trips with missing or different address	1,664	17.2
Trips with missing personal info for mode choice model	458	4.7
Trips without any missing or different info for mode choice model	**7,087**	**73.2**

Activity data:

Table 3 shows that 3.8% of the time is not reported at all and 2.2 % were reported but with missing activity type in the pre-validation data. This results mainly in under-reporting of sleeping (−3.2 %), and minor changes in leisure (−0.8 %) as well as personal and domestic activities (−0.4 % and −0.5 %). The full paper will include a more detailed analysis of the characteristics of missed activities.

Table 3: Activities by activity type in pre- and post-validation data

Activity type	post-validation min/day	post-validation share [%]	pre-validation share [%]	difference [%]
missing	0	0.0	2.2	*2.2*
travel	86	6.0	5.9	*-0.1*
sleep	482	33.5	30.3	*-3.2*
eating	81	5.6	5.2	*-0.4*
work	295	20.5	20.0	*-0.4*
education	11	0.7	0.7	*0.0*
personal	89	6.2	5.7	*-0.4*
domestic	111	7.7	7.2	*-0.5*
shopping	18	1.2	1.1	*-0.1*
leisure	263	18.3	17.5	*-0.8*
other	3	0.2	0.3	*0.0*
unspecific	0	0.0	0.0	*0.0*
Total	**1,440**	**100.0**	**96.2**	

Effort of validation:

Table 4 gives an overview of the efforts spent for different parts of the field work. The validation took on average 15.6 minutes per respondent. This amounts to 12 % of total effort of all survey team activities within the survey procedure. It should be noted that the validation effort also includes the validation of expenditure data, which is not part of this study.

Table 4: Shares of different activities of the survey team

Kind of activity	min per respondent
Motivation	25.4
Support & reminders	12.0
Data entry	80.1
Validation	**15.6**
Others	0.1
Total	**133.1**

5 Conclusions and outlook on further contents of the full hEART-paper

In the validation process, only few additional items of trips and activities emerged, because the MAED survey provides a fairly complete dataset from the outset (unlike conventional travel diaries). Travel data required much validation due to missing key trip information (mainly addresses) and key personal information (mainly mobility tools). Validation was successful in gathering this information. Activity data were in most cases correctly reported from the outset and required little validation. Validation took 12 % of total effort in the field work process. Some effort could be saved in an online survey when e.g. addresses are reported more easily and correctly by clicking on a map, when answers on specific items are forced or when prompts and reminders support respondents in filling out the questionnaire.

6 Literature

Aschauer, F., Hössinger, R., Axhausen, K. W., Schmid, B., Gerike, R. (2018). Implications of survey methods on travel and non-travel activities: A comparison of the Austrian national travel survey and an innovative mobility-activity-expenditure diary (MAED). European Journal of Transport and Infrastructure Research, 18 (1), 4-35.

Brög, W., Erl E., Meyburg, A., Wermuth. M. (1982). Problems of nonreported trips in surveys of nonhome activity patterns. Transportation Research Record. 891, 1-5.

Gerike, R., Gehlert, T., Leisch, F. (2015). Time Use in Travel Surveys and Time Use Surveys – Two sides of the same coin? Transportation Research Part A, Special Issue on Time Use. 76, 4-24.

Richardson, A. J., Ampt, E. S., Meyburg, A. H. (1995). Survey Methods for Transport Planning. Eucalyptus Press, Melbourne.