CONSTRUCTION OF TREE PANEL WEIGHTS

Documentation for the panel waves from 2000 to 2010

STEFAN SACCHI

TREE & cue sozialforschung, Basel/Zurich, 2011
www.tree.unibas.ch

Contents

Introduction

The longitudinal youth survey 'Transition from Education to Employment' (TREE) is based on a panel survey focusing on youths' pathways from school to working life. The Swiss section of the *PISA study* conducted in spring 2000 (PISA, 2002) served as the *first wave* of the panel survey. This internationally comparative survey collects data on basic competencies, such as reading proficiency, basic knowledge in mathematics and the natural sciences, along with detailed information on important background and context factors. The TREE project follows up those youths initially surveyed and tested by PISA in 2000. Seven survey panels have been conducted on a yearly basis from 2001 to 2007, an eighth panel in 2010. This longitudinal design is geared to obtaining representative longitudinal information about the difficulties encountered – and how they are dealt with – in the transition from school to vocational education and training, and later on to the labour market.

This documentation describes the longitudinal sample weights for the initial PISA survey and the first eight TREE panel waves. It is an update of the 2008 documentation for the first seven TREE panel waves (Sacchi 2008a, 2008b), which in turn was based on previous working papers (Sacchi, 2003, 2004a, 2004b).

1 Sample and population

The PISA sample serving as the basis for the TREE panel is designed to be representative for ninth graders as well as for fifteen-year-olds irrespective of the grade they had attended at the time. It involves a two-stage, multiply disproportionate random sample based on predefined sample sizes for the two groups just mentioned, for the language regions, and the participating cantons (for details see Renaud, Ramseier & Zahner, 2000; PISA, 2002). Apart from this, an independent class sample was drawn from all ninth grade classes in French-speaking Switzerland, and all students from each of the classes selected were surveyed (single-stage cluster sample, cf. PISA Romandie, no date).

The TREE population is defined as the subset of PISA respondents who at the time of the PISA survey had attended a regular public school anywhere in Switzerland or a regular private school in Italian-speaking Switzerland and had not yet completed compulsory education at the time but then left compulsory school at the end of the 1999-2000 school year. The TREE study population is therefore for the most part the same as the subsample of ninth graders in PISA.[1]

The exception that needs to be mentioned is that students of private schools have been included only in Italian-speaking Switzerland and not in the two other language regions.[2] Moreover, PISA respondents who had not yet left compulsory school one year *after* the initial PISA survey were not included in the TREE population.

---

[1] Additionally included is a small group of youths from the sample of 15-year-olds who attended seventh or eighth grade at the time of the PISA survey and who prematurely dropped out of compulsory education during the 1999/2000 school year (≈ 1 % of the initial TREE sample).

[2] This reduces the PISA sample (N=13467) by 673 respondents or approximately 5 percent.

2 Survey process and response

For reasons of data protection, the TREE panel surveys required that consent to participate already be obtained from the student target group during the PISA survey. For this purpose, a target group specific information sheet about the TREE project was distributed during the PISA test sessions asking PISA participants to also participate in TREE. In addition, a special information leaflet was handed out to inform PISA test administrators (cf. Meyer, 2000). Youths willing to participate in TREE were asked to return the information sheet containing their address information. An explorative analysis of address return rates clearly shows signs of *regional or local test administration*, as the case may be, representing a crucial factor affecting participation in the address survey (Meyer, 2000: 4). Those youths who volunteered their addresses were surveyed in spring 2001 (wave 1) and then again every spring in yearly intervals. *Table 1* gives an overview of the survey process up to the seventh and so far last follow-up survey in 2010.

It turned out only in retrospect that 727 cases out of the 7,070 PISA respondents who provided their addresses for the panel study did not fit the criteria for inclusion in the study population.[1] The large majority of those youths (N=608) were either attending compulsory school or repeating ninth grade at the time of the first TREE panel wave in 2001. Another good 100 cases were excluded because they had not been attending a type of school included under the TREE sampling criteria at the time of the PISA survey (particularly special schools were excluded). The remaining sample for the first TREE panel wave thus comprises 6343 cases. *Table 1* documents the initial TREE sample, cumulative sample attrition across panel waves (in this respect also note the information provided at the bottom of the table), and the response rates of the seven panel waves based on the revised initial sample. An excellent response rate of between 76 and just under 88 percent was achieved in each wave. Cumulatively, this resulted in a still sizeable participation rate of close to 54 percent up to the eighth wave. More than 41 per cent of the youths in the initial sample participated in *all* of the eight panel waves. This is a very good result for such a long, multi-wave panel survey.

The exceptionally high level of overall participation has been achieved by offering potential dropouts alternative modes of participation. For instance, in the first four TREE waves, participating youths were given the option of responding to the survey questions by phone instead of completing the self-administered questionnaire, and – if necessary – a considerably shorter instrument was used (short CATI interviews). From the fifth wave on, TREE has employed a combination of telephone interviewing and a written questionnaire as the standard mode of administering the survey. This *change in methods* was in order because of the increasing diversity of individual education and employment paths as respondents grew older, which would have required increasingly complex sequences of filter questions to the point of rendering a written questionnaire unmanageable. Bearing in mind the need for maintaining intraindividual comparability over time, it is still important to obtain information underlying a number of items and psychological scales prone to mode effects (Klein & Porst, 2000; Scherpenzeel, 2001) in written form.

[1] The initial TREE sample was singled out mostly based on information obtained in the first wave and, in the second instance, based on indirect information about non-respondents (from contact records).

Table 1: *Initial sample, sample realized, and panel attrition*

(Absolute numbers)	National PISA sample	Class sample French-speaking Switzer-land	Combined initial sample	Sample realized (n)	Response rate (%)
1. PISA survey					
Initial PISA sample	10,423	?	?		
participants outside of PISA population [1)]	101	?	?		
absent at time of survey	150	?	?		
PISA survey	10,172	5,073 [2)]	15,241	14,494	95.1 %
participants included in both PISA samples				– 1,031	
Composite PISA sample				13,463	
participants outside of TREE population [3)]				– 673	
Initial sample for address survey				12,794	
2. TREE panel					
Address survey			12,794	7,070	55.3 %
not included in the TREE population [4)]				– 727	
Initial TREE sample				**6,343**	
Panel wave 1			6,343	5,532	87.2 %
final drop out by wave 1			– 400		
Panel wave 2			5,943	5,210	87.7 %
final drop out by wave 2			– 344		
Panel wave 3			5,599	4,880	87.2 %
final drop out by wave 3			– 266		
Panel wave 4			5,333	4,680	87.8 %
final drop out by wave 4			– 284		
Panel wave 5			5,049	4,504	89.2 %
final drop out by wave 5			– 205		
Panel wave 6			4,844	4,135	85.4 %
final drop out by wave 6			– 204		
Panel wave 7			4,640	3,982	85.8 %
final drop out by wave 7			– 120		
mistakenly not contacted in wave 8			– 15		
Panel wave 8			4,505	3,424	76.0%
Cumulative participation TREE T1 – T8				3,982	54.0 %
number participating in *all* 8 waves				2,618	41.3 %

1) Unable to complete PISA test session.
2) Indirectly inferred from sample weights.
3) Not included in the TREE population were youths attending private schools (with the exception of Italian-speaking regions), 15-year olds in non-compulsory education, youths from Bernese Jura.

Two non-response surveys that followed up on the TREE waves 2003 and 2004 also suggest switching data collection methods (Stalder & Dellenbach; 2005). The two follow-up surveys each asked more than 1000 youths about their reasons for either refusing to participate in the respective panel wave at all or only participating by phone. The non-response analyses identify time constraints as the main reason given for not participating in the written survey or not

participating at all. Moreover, youths opting for the telephone survey (short or long version) criticize the length and increasing complexity of the written questionnaire. The reasons offered for complete refusal, however, tend to be a lack of interest in the topic and reluctance to continue regular participation in the study. This feedback supports arguments in favour of switching to a method mix focusing on CATI interviews that should be clearly shorter than the longer optional telephone interview.

Hence, there are good theoretical and empirical reasons calling for the shift implemented in the *fifth wave* to a mixed method design combining a fairly short telephone interview and a written supplementary questionnaire. From wave 5 on, respondents in addition to taking the survey in the standard mixed form have been given the choice of either taking the long version entirely in written form or responding only to a reduced set of questions either by phone *or* in written form (mostly CATI, in individual cases in written form).

As shown in Table 2, the proportion of respondents willing to participate only in the shorter survey has risen sharply from the third wave on. Initially hovering around two to three percent, from T3 on that share climbs to fifteen percent and reaches the 20 percent mark by the fifth wave. The option of taking the shorter survey (Optional Mode 2) apparently has been a major factor in accounting for the pleasingly high participation rate over time. In addition, the standard use of a mixed mode design has probably also been a factor since it has significantly facilitated the participation of youths who are less proficient in dealing with written texts.

Table 2: *Type of participation by panel wave*

Survey method (share of respondents)	Standard Mode	Optional Mode 1: full list of questions	Optional Mode 2: short list of questions
TREE panel wave			
Wave 1 (N = 5,532)	Questionnaire (91.3 %)	CATI long (6.6 %)	CATI short (2.2 %)
Wave 2 (N = 5,210)	Questionnaire (91.7 %)	CATI long (5.0 %)	CATI short (3.3 %)
Wave 3 (N = 4,880)	Questionnaire (81.7 %)	CATI long (3.0 %)	CATI short (15.4 %)
Wave 4 (N = 4,680)	Questionnaire (81.3 %)	CATI long (5.2 %)	CATI short (13.5 %)
Wave 5 (N = 4,504)	CATI plus written (76.5 %) [1)]	Questionnaire (3.1 %) [2)]	CATI (20.5 %) [3)]
Wave 6 (N = 4,135)	CATI plus written (79.7 %) [1)]	Questionnaire (1.1 %) [2)]	CATI (19.3 %) [3)]
Wave 7 (N = 3,982)	CATI plus written (74.0 %) [1)]	Questionnaire (5.7 %) [2)]	CATI (20.3 %) [3)]
Wave 8 (N = 3,424)	CATI plus written (71.6 %) [1)]	Questionnaire (9.1 %) [2)]	CATI (19.3 %) [3)]

1) Supplementary questionnaire especially containing various psychological scales prone to mode effects. Share includes cases who broke off the CATI interview.
2) Basic written (instead of CATI interview) plus supplementary questionnaire.
3) Supplementary interview not completed; includes cases who completed basic written questionnaire instead of CATI.

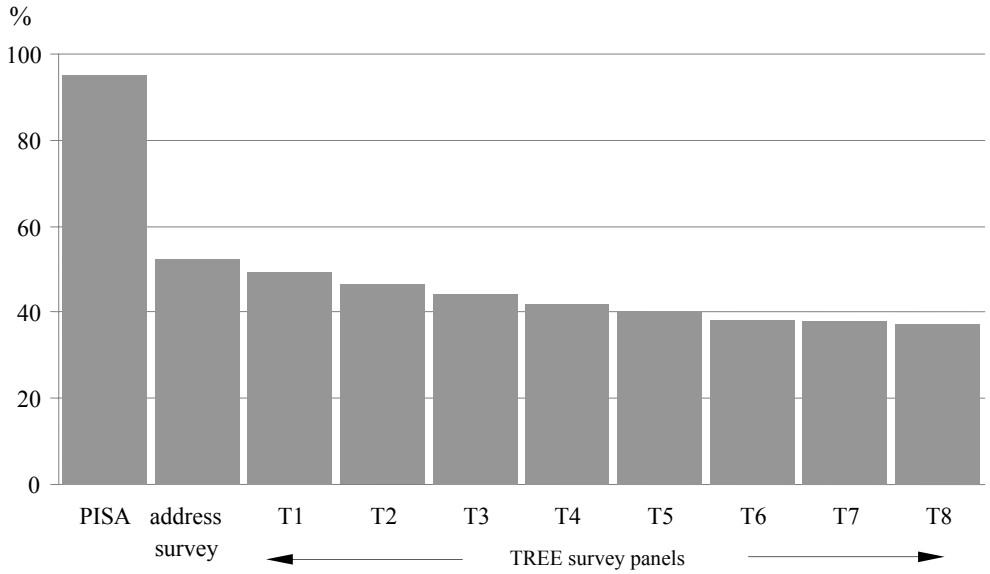
Table 1 already shows that a significant share of sample attrition occurred prior to the first TREE survey as the address survey was conducted in connection with administering the PISA test. This is illustrated in the following diagram, which reflects the gradual decline in the size of the remaining panel sample across the individual waves.[1]

Since sample attrition and non-response only in rare cases can be expected to occur at random (Schnell, 1997), panel weights are usually used to compensate for the potential bias in sample composition (see for instance the comparison of methods by Rizzo, Kalton & Brick, 1994).

[1] Sample attrition shown in the diagram refers to cases of *final* drop out from the sample.

As far as TREE is concerned, the main effort should be devoted to non-response in connection with the address survey by PISA test administrators prior to the actual TREE survey since non-response by far has occurred most frequently at this stage (see Diagram 1).

Diagram 1: *Cumulative impact of panel attrition (initial PISA sample = 100%)*[1]



Compared to cross-sectional surveys, it is far easier to correct for non-response bias in subsequent waves of a panel survey, as much more extensive information about non-respondents is available from previous waves. Regarding the TREE weights, it is very fortunate that, at approximately 5 percent, non-response was very low in the PISA survey that served as the first wave of the TREE panel. All information about respondents and the conditions surrounding the survey obtained in the course of PISA are thus available for correcting for the high level of non-response encountered in the address survey. In addition, all the information from any given TREE wave can be used to correct for non-response occurring in subsequent waves.

[1] Share of respondents who remain in the sample up to the respective panel wave (i.e. who do not permanently withdraw from participating in the survey).

3 Construction of the TREE panel weights

Panel weights for samples of individuals are usually constructed as the reciprocal of the product of the individual response probabilities for the different panel waves (cf. Sacchi 2001) – the German Socio-Economic Panel is a case in point (Haisken-DeNew & Frick, 2000, 140 f.). For the TREE panel the following relation thus holds:

$$G_i = \frac{1}{E_{PISA.i} \cdot A_{PISA.i}} \cdot \frac{1}{A_{ADR.i}} \cdot \frac{1}{A_{W1.i}} \cdot \ldots \cdot \frac{1}{A_{Wt.i}} \qquad (1)$$

where

G_i	longitudinal raw weight for panel wave *t* for respondent *i*
$E_{PISA.i}$	probability of inclusion of *i* in the initial PISA sample
$A_{PISA.i}$	probability of participation of *i* in the PISA survey
$A_{ADR.i}$	probability of participation of *i* in the TREE address survey
$A_{W1.i}$	probability of participation of *i* in TREE panel wave 1
$A_{Wt.i}$	probability of participation of *i* in TREE panel wave *t*

The probabilities of participation $A_{..i}$ are conditional probabilities, thus referring to the likelihood of response provided that a respondent *i* is part of the initial PISA sample and does not drop out of the panel sample in one of the previous waves. The quantitatively most significant source of panel attrition is *final* refusal to participate not just in the wave in question but also in *all* future waves (see notes at the bottom of Table 1).[1] In addition, in each wave there are a handful of youths who are deleted from the survey population due to migration, decease or relocation and who could not be contacted.

With regard to the TREE panel, a weighting model as defined by equation (1) has the following advantages:

- The model builds on existing PISA weights, which correct for design effects and non-response pertaining to the PISA sample. The weights are constructed as the reciprocal of the product of $E_{PISA.i}$ and $A_{PISA.i}$ as defined by equation (1). As is common practice in panel weight design, the other components of G_i are determined using logit models or logistic regression. This allows *systematically* taking into account interindividual differences in the probabilities of participation.
- Although perhaps not by all means necessary, separation of $A_{ADR.i}$ and $A_{W1.i}$, on the one hand, has the advantage of allowing to focus on estimating $A_{ADR.i}$ in the process of model development; since non-participation in the address survey is the main source of non-response, a good approximation of $A_{ADR.i}$ is crucial in correcting for potentially ensuing bias. On the other hand, this measure takes into consideration that to some extent different factors can be expected to be critical in determining $A_{ADR.i}$ and $A_{W1.i}$ respectively: It has been pointed out that the situational context surrounding the PISA test sessions greatly influences participation in the initial collection of address data whereas participation in subsequent panel waves depends more on individual attributes.

[1] Refusal to participate in any one wave does not lead to deletion from the panel.

- Owing to a modular design, weights for further panel waves and the corresponding adjustments for non-response can easily be added if necessary by multiplying G_i, as defined by equation (1), with the reciprocal of the response probabilities of the additional wave(s).

3.1 Design weights for the composite PISA sample

As mentioned above, the Swiss PISA survey consists of two *independent* random samples. For each of the samples there exists a specific weight variable designed to compensate for unequal selection probabilities ensuing from sample design, and also for non-response. Non-response adjustments correct for non-participation of some schools (national sample) and classes respectively (class sample French-speaking Switzerland) (also see Table 1). As is common in cross-sectional surveys, the non-response corrections are rudimentary and not without problems yet of marginal significance in this context since non-response lies at only about five percent.[1] The construction of both PISA weights is documented elsewhere (Renaud, Ramseier & Zahner, 2000; PISA Consortium, 2000; PISA Romandie, no date). The important point here is that for both samples the weights equal the reciprocal of the individual probability of inclusion as required by equation (1).

However, calculation of TREE panel weights as defined by equation (1) requires a weight that corrects for design effects and non-response for the *composite* PISA sample including both the national PISA sample and the independent class sample for French-speaking Switzerland. In case of German and Italian-speaking Switzerland, where only one sample was drawn, the weight variable of the national PISA sample (*'w_fstuwt'*) can be readily employed for this purpose. Conversely, the weight variable for the class sample can be used in case of the canton of Jura, where an exhaustive sample was taken.[2]

For the remaining parts of French-speaking Switzerland, a new design weight must be constructed to adjust for the fact that youths from that region have two independent chances of being selected. Since we are dealing with two independent samples, we may calculate the inclusion probability of those youths according to the addition law of probabilities as follows:

$$P_{Rom.i} = P_{N.i} + P_{C.i} - \left(P_{N.i} \cdot P_{C.i}\right) \tag{2}$$

where

$P_{Rom.i}$	probability of selection of respondent *i* from French-speaking Switzerland
$P_{N.i}$	probability of inclusion of *i* in the national sample
$P_{C.i}$	probability of inclusion of *i* in the class sample from French-speaking Switzerland

In principle, the probability of being selected and the PISA design weight for French-speaking Switzerland as its reciprocal can be easily calculated by entering the reciprocals of the two PISA weight variables for the national sample and class sample for $P_{N.i}$ and $P_{C.i}$ respectively in equation (2). The resulting design weight thus already factors in the adjustments for non-response contained in the two PISA weights.

---

1 PISA non-response adjustments are based on the somewhat questionable assumption that the respondents from one school (national sample) or one class (class sample French-speaking Switzerland) respectively are also representative of the non-respondents.

2 Because of adjustment for non-response, the weights used for the canton of Jura are not constant and for the most part slightly greater than 1 in spite of the exhaustive sample.

In practice, however, we face the problem that *both* weight variables are only available in cases in which respondents happen to be included in both samples. This, however, is the case for only 309 out of 4930 youths of the composite sample drawn from French-speaking Switzerland (excluding Jura). By contrast, for 3806 youths only the weight for the class sample is available and for 828 youths solely the one for the national sample. Hence, for those two groups the missing weights must be reconstructed.

The task is easily accomplished as far as *reconstruction of weights for the class sample* is concerned, as it is a single-stage cluster sample with a fairly simple weighting scheme. Calculation can start from the reasonable assumption that stratum-specific non-response rates among the 828 respondents outside of the class sample would be the same as those observed within the individual strata of that sample. The potential impact of that assumption is further limited for the fact that the adjustment for non-response accounts for only approximately four percent of PISA weight variance in the class sample from French-speaking Switzerland.[1]

On the other hand, the *reconstruction of weights for the national PISA sample* has proven impossible based on the available data. For a retrospective estimate one would definitely have to know the total number of students belonging to the PISA population for all schools included in the class sample (Renaud, Ramseier & Zahner, 2000, 8; PISA Consortium, 2000, 7). This information is available for schools represented in the national sample but not for those that are part of the class sample only. Moreover, attempts at reconstructing this information in retrospect have failed just as have efforts at approximating the national weights based on the data at hand.

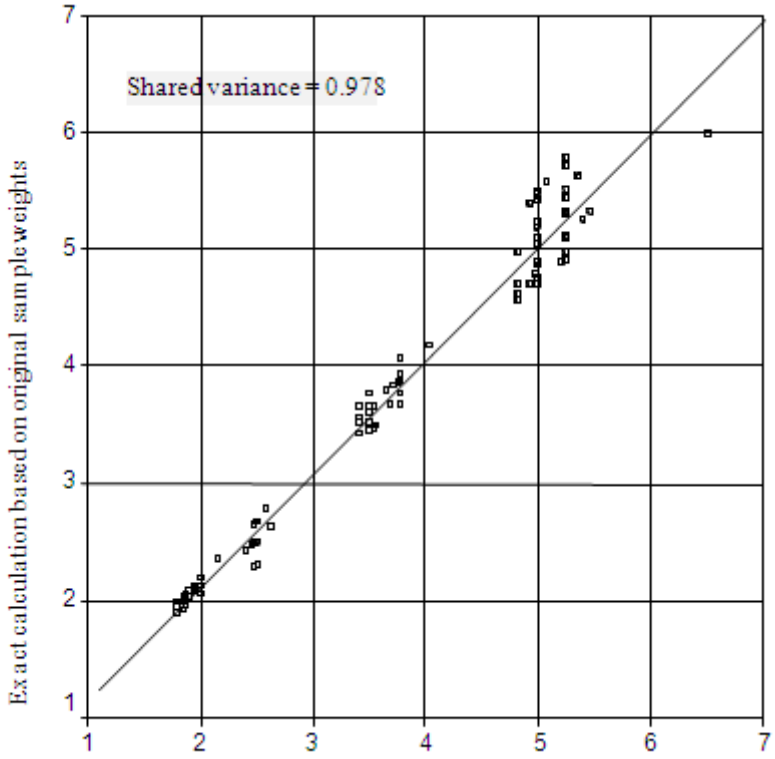
The only feasible solution is to approximate the weight variable for the national sample of those 3806 youths of the class sample by substituting the mean of the national weight variables for all ninth graders from French-speaking Switzerland (excluding Jura, only stratum 23) for the missing weight.[2] This approximation implies that the probability of inclusion into the national sample is the same for all of the 3806 youths. Although this is a rather unsatisfactory solution from a theoretical point of view, the practical effects in terms of the quality of the resulting sample weight are small. At any rate, in the case of those ninth graders from French-speaking Switzerland who belong to the national sample and for whom the original sample weights are known, the design weights as defined by equation (2) remain roughly unchanged when substituting the actual national weight variable by the said mean value. Calculations with and without mean substitution are illustrated in the following diagram.

---

1 This is the result of a variance analysis where the stratum variable defining the subpopulations with invariant probabilities of selection serves as the factor and the weight as the dependent variable.

2 The mean of the weight variables is 13.9.

Diagram 2: *Calculation of design weights for French-speaking Switzerland. Assessing the loss of precision due to mean substitution (N=1034)*



The extremely high correspondence irrespective of neglecting the significant differences in the individual probabilities of being included in the national sample results from sampling fractions of the class sample being several times greater. For this reason, the class-sample weights affect the design weights, calculated as the reciprocal of the inclusion probabilities according to equation (2), to a much greater degree than the national weight variable. Thus, the inevitable substitution of the approximately 3800 missing values in the national weight variable by the mean fortunately has no substantial effect on the quality of the design weight.

To sum up, the design weight may be defined as follows:

$$G_{PISA.i} = \begin{cases} G_{Rom.i} & \text{stratum 22 (canton of Jura)} \\ \left(1/G_{Rom.i} + 1/G_{Nat.i} - \left[1/G_{Rom.i} \cdot 1/G_{Nat.i} \right] \right)^{-1} & \text{stratum 23/61f. (remaining Romandie)} \\ G_{Nat.i} & \text{remaining strata (remaining Switzerland)} \end{cases} \quad (3)$$

As mentioned before, the class sample weight ($G_{Rom.i}$) can be used as design weight for the *canton of Jura* as an exhaustive sample was taken there. The design weight for remaining French-speaking Switzerland is calculated according to equation (2) while the weight variable for the national PISA sample ($G_{Nat.i}$) is entered in the case of German and Italian-speaking Switzerland.

3.2 Estimation of response probabilities

The probabilities of participating in the address survey as well as in the seven yearly TREE panel waves since 2001 are estimated using logistic regression (Hosmer & Lemeshow, 1989) where the information about family and school background as well as individual attributes collected in previous PISA or TREE waves may serve as predictors. Following a similar philosophy as Wießner (2003: 89), the specification of the models to correct for non-response bias are based as much as possible on the arguments and findings of non-response research (Schnell 1997; Koch & Porst, 1998; Stoop, 2005); yet, in lack of a sophisticated theory of participation, behaviour model construction is to some extent forced to rely on inductive reasoning. All attributes of the youths, their families and educational backgrounds that might plausibly be expected to have an influence on participation behaviour are thus tentatively included in the participation models. In modelling address survey response, the school setting and other situational factors that might have had an effect during the PISA test sessions are also considered. To the extent that the factors identified have proven theoretically plausible, statistically significant ($\alpha \leq 1\%$) and robust in terms of effects on participation, they are included in the definite models for the respective panel wave.[1] Thus, model construction inevitably involves an inductive approach to some degree. Proceeding in this way, on the one hand, has the advantage of fairly comprehensively taking potentially significant predictors of non-response into account, thus minimizing *omitted variable bias* (Menard 2002: 68f.). On the other hand, there is a risk of *overfitting* the model to random distributions idiosyncratic to the particular sample. In this light, it is always good advice to assess the theoretical plausibility of any changes in sample estimates caused by the resulting non-response adjustments (also see section 7).

The participation probability estimates are calculated based on the effect coefficients and the individual values for the variables as defined by the following equation (cf. Menard, 2002: 13) and entered into equation (1):

$$A_i = \frac{e^{\left[B_0 + \sum_{j=1}^{J} B_j \cdot X_{ji}\right]}}{1 + e^{\left[B_0 + \sum_{j=1}^{J} B_j \cdot X_{ji}\right]}} \tag{4}$$

where

A_i	estimated probability of participation of respondent *i*
B_0	regression constant
B_j	effect coefficient for variable *j*
X_{ji}	value of variable *j* for respondent *i*

The individual models for participation in the address survey and the seven TREE panel waves are described in detail in the German documentation of panel weight construction (especially Tables 3 to 10). The following synopsis is limited to the most significant effects of non-response on panel composition that have gradually emerged in the course of the seven TREE panel waves.

[1] Because of the sizeable sample, a significance level (α) of 1 percent seems appropriate. The models are checked for robustness by tentatively excluding the most influential cases in terms of cook distances from estimation.

3.3 Cumulative effects of non-response

Before we turn to truncation, poststratification and calibration of the raw weights as defined by equation (1) in the sections to come, I shall briefly discuss the cumulative effects of non-response on the composition of the remaining panel after each wave.

In principle, we can expect the differences in sample inclusion probabilities for any given wave to be partly related to the disproportional design of the initial PISA sample and partly to variations in individual participation behaviour. In fact, the relative impact of participation behaviour increases over consecutive waves compared to the differences stemming from sample design. Accordingly, the proportion of purely design-related differences in the individual probabilities of panel inclusion (reciprocal of weights as defined by equation 1) drops from 61 to 37 percent by the eighth wave. The variance resulting from sample design alone is of no concern in this context because it is free of bias arising from specification or estimation errors. The models correcting for systematic differences in participation behaviour, on the other hand, are at best good approximations of the underlying (self-)selection processes. From this vantage point, a much welcome circumstance is that a major – yet with each wave smaller – portion of the differences in individual inclusion probabilities can be traced to the grossly disproportionate PISA survey design. In other words, variation in inclusion probabilities is mostly accounted for by large *differences in the sampling fractions* among the different strata of the composite PISA sample. Under the bottom line, the design weights for the most part simply correct for the strong overrepresentation of French-speaking Switzerland in the initial sample.

Regarding the *structure* of non-response, the first issue calling for attention is how the relationship between the attributes at the centre of TREE research and respondents' willingness to participate in the survey has changed across the panel waves. Those attributes in the narrow sense include all individual attributes excluding indicators relating to willingness to respond (participation behaviour thus far), test administration and context variables. The following table contains two different fit statistics for each wave, namely a McFadden-R^2 for the complete model for each wave and one for a corresponding model *without* individual attributes. Comparison of the two values reveals the contribution of individual attributes in the narrow sense to model fit.

Especially in case of the first three and again the last wave, model fit seems to crucially depend on the individual attributes included. Thus, in the case of those waves in particular, there is a quite strong relation between relevant individual attributes and non-response. However, as far as the first three waves are concerned, the *overall* relationship between predictors and participation is a fairly modest one, as the McFadden-R^2 for the complete model indicates. This can be taken as a sign that, although individual attributes strongly contribute to model fit, their impact on non-response is quite limited. In the models for waves 4 to 6, individual attributes apparently play only a marginal role. Since the TREE attributes have been scrutinized quite extensively for potentially significant predictors of participation, these findings altogether suggest that differences in non-response for the most part stem from sources that are unrelated to the attributes under study. Considering the potential for non-response bias, this is of course a welcome finding. Participation in the seventh panel wave, however, depends to a larger extent on various relevant individual attributes (Table 10), resulting in a potentially much stronger bias. In the eighth TREE wave, there fortunately seems to have again been only very little change in the composition of the panel sample.

Table 3: *The contribution of individual attributes to model fit by wave*

Response model	Likelihood ratio pseudo R^2 (McFadden)	
	Complete model	Reduced model [1)]
Address sheet	.202	.140
Wave 1	.096	.032
Wave 2	.130	.001
Wave 3	.131	.048
Wave 4	.185	.161
Wave 5	.154	.122
Wave 6	.207	.190
Wave 7	.391	.254
Wave 8	.210	.204

1) Models without individual attributes and related interactions.

However, close inspection of the wave-specific participation models reveals that a number of individual attributes affect participation across several waves resulting in a cumulation of effects on sample composition over time. To assess such *cumulative effects* on panel composition, all individual attributes showing a substantial impact on participation over several waves are considered as relevant. The conditional probability that youths with such attributes remain in the panel up to a given wave is first calculated based on the wave-specific models of participation and then divided by the corresponding probability for 'average' youth without the attribute in question.[1] The resulting *'relative inclusion probabilities'* in Table 4 thus give an idea of how much the probability of remaining in the panel depends on the most pertinent individual attributes and how it changes across waves.

The first two rows of the table, for instance, illustrate that youths achieving very high levels of *reading proficiency* much more frequently remain in the panel and those exhibiting lower proficiency scores more often drop out compared to 'average youths'. Already in the address survey, the probability of the first group remaining in the initial sample is about 24 percent above average while the percentage for the second group is below average almost by the same amount. By the fifth wave, panel composition in terms of reading proficiency has gradually become even more skewed to the point that the probability of remaining in the sample for the two groups in the last three waves is at 53 percent above and 42 percent below average respectively. The parallel effects of reading proficiency across the first waves thus cumulate to a considerable degree. The following diagram illustrates the results for the five PISA proficiency levels (PISA 2002: 24f.).

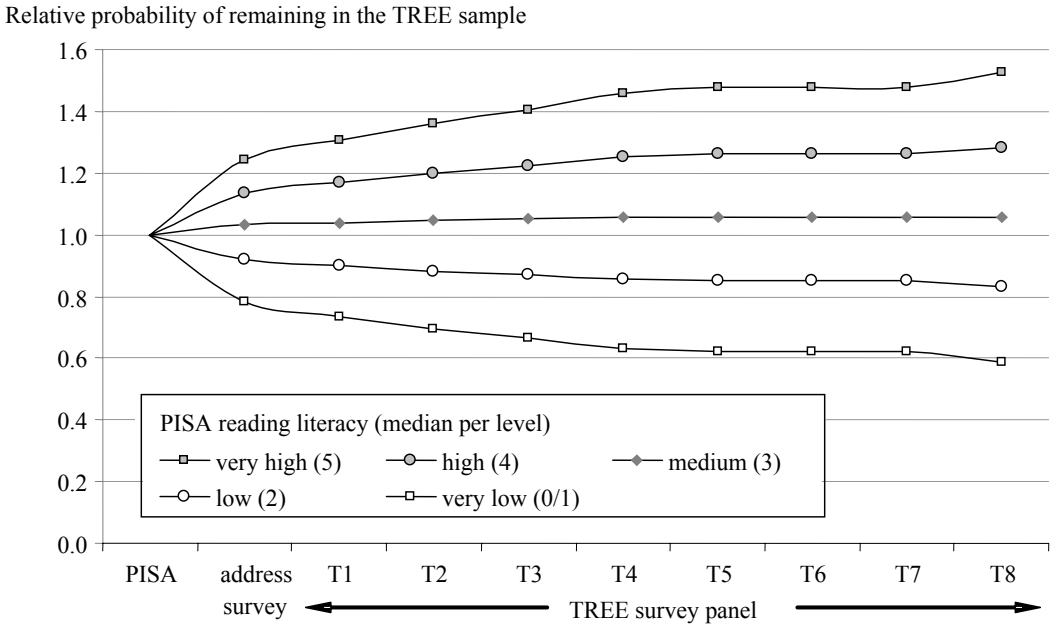
[1] Based on model estimates and equation (4), wave-specific participation probabilities are calculated and then multiplied out across waves to arrive at the conditional probability of remaining in the panel up to wave x. When calculating the wave-specific probabilities for youth with and without a given attribute, all the other predictors are set to their mean (scales, ratings) or modal values (categorical variables).

Table 4: *Cumulative effects of selected predictors on panel attrition**

Relative probability of remaining in the sample [1]		TREE panel wave								
Attributes relevant to participation [2]		Addr. [4]	T1	T2 [5]	T3	T4	T5	T6	T7	T8
PISA reading proficiency [3]	very high	1.24	1.31	1.36	1.40	1.46	1.48	1.48	1.48	1.53
	very low	0.78	0.73	0.69	0.67	0.63	0.62	0.62	0.62	0.58
Not in ninth grade		1.20	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Plans for the future: continue vocational education and training (VET)		1.00	1.00	1.00	1.00	1.05	1.08	1.06	1.08	1.11
Plans for the future: other/different VET programme		1.00	1.00	1.00	1.00	1.05	1.07	1.07	1.09	1.09
VET experience: not as expected		1.00	1.00	1.00	1.03	1.03	1.05	1.05	1.05	1.05
Homework completed on time: never		1.00	0.92	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Score on addictive substance consumption scale [3]	very high	1.00	1.00	0.98	0.96	0.96	0.96	0.96	0.96	0.96
Score on the self-confidence scale	very low	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.92
At home: number of mobile phones (≥3)		1.00	0.97	0.95	0.95	0.92	0.92	0.92	0.92	0.92
At home: number of calculators (≥3)		1.00	1.04	1.07	1.07	1.07	1.07	1.07	1.07	1.07
Does not live with mother		1.00	0.92	0.89	0.89	0.89	0.89	0.89	0.89	0.89
Does not live with father		1.00	0.96	0.93	0.93	0.93	0.93	0.93	0.90	0.90
Born outside of Central Europe		0.92	0.86	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Gender: female		1.16	1.21	1.25	1.28	1.28	1.28	1.28	1.25	1.25

*Minor errors in the 2008 version have been corrected.
1) Relationship between the probability of a youth with a certain attribute remaining in the panel sample and the corresponding probability of an 'average' youth doing so (see text).
2) Attributes that have proven relevant to models for at least *two* of the panel waves.
3) 'Very high' and 'very low' respectively equals the median of youths achieving the PISA reading proficiency level five and one or less, respectively, in the initial TREE sample (wleread = 650 and 359 respectively).
4) In calculating the probability of participation in the address survey, an average interaction effect between reading proficiency and test administration is assumed (cluster 9 according to Table 3).
5) In calculating the probability of participation in wave 2, 'homework always completed | missing' serves as a reference category for the respective variable.

Diagram 3: *Remaining in the TREE panel by PISA reading literacy level*



The diagram, on the one hand, demonstrates that from the fifth wave on the probability of remaining in the panel sample is two and a half times higher for youths with very high reading proficiency (PISA level 5) compared to youths with very low reading skills (level ≤ 1).[1] Across the panel waves, reading proficiency thus has a huge impact on panel attrition. On the other hand, it clearly shows that non-participation in the address survey already accounts for about half of the difference in the relative inclusion probabilities right at the beginning. Those initial differences steadily increase with each additional wave and remain at a constantly high level from the fifth wave on. Switching to CATI as the basic TREE module (see Table 2) from the fifth wave on may have contributed to this welcome stabilization.

Table 4 documents the findings for all attributes that play a role in at least two of the wave-specific participation models and therefore can *potentially* be expected to entail cumulative bias with respect to the composition of the panel sample. In contrast to reading proficiency, however, the cumulative effects in these cases, with few exceptions, are fairly small. For the majority of the attributes, the relative inclusion probabilities in Table 4 range from 0.9 to 1.1 across all waves, implying individual deviations of approximately 10 percent at maximum from the average.

Greater deviation, however, is observed for country of origin and gender at the bottom of the table, *on the one hand*. For *youths of foreign origin,* the large majority of whom come from other than Central European countries, the probability of remaining in the panel drops to 83 percent of the average probability by wave 2. From then on it remains constant, ceteris paribus. *Young women*, by contrast, already display a much higher level of participation in the address survey (+ 16 %) and this overrepresentation in the panel climbs to 28 percent by the third wave. After remaining constant for four waves (T3-T6), the effect has somewhat declined in the past two waves, (+ 25%).

On the other hand, it needs to be mentioned that the cumulated effect of not living with mother and/or father at an early stage, which each reduces the inclusion probability by approximately ten percent, goes hand in hand with a number of wave-specific effects resulting from family situation, residential environment and critical life events that point in the same direction. Sharing an apartment with others (T2), cohabitation (T4), leaving the parental home at an early stage (T3, T6) and early parenthood (T5) all exert a negative effect on continued participation in the panel. Overall, these findings suggest that youths from incomplete families as well as those who leave home early are more strongly underrepresented in the panel than Table 4 alone would lead us to expect.

Summing up, we may conclude that the composition of the sample in the course of the panel waves has changed primarily with regard to four individual attributes. On the one hand, youths with low reading proficiency, young men and youths of foreign origin had already dropped out of the sample at an above average rate at the time of the address survey. The first panel waves subsequently reinforced their already sizable underrepresentation in the initial TREE sample. On the other hand, youths from complete families who had been living in the parental home during the whole period under study are markedly overrepresented in the panel sample.

[1] From wave 5 on, the ratio of the respective probabilities of remaining in the sample stays at a constant level of about 2.4 (1.48 / 0.62).

4 Truncation of the raw weights

After estimating participation probabilities for all waves, the elements needed for calculating the raw weight as defined by equation (2) are available. Under the assumption that the wave-specific models accurately depict the differences in individual participation probabilities, these panel weights allow an unbiased estimation of population characteristics.

Yet, in applying such sample weights, there often exists a *trade-off* between *correcting for non-response bias* and minimizing the adverse impact of weights on the *accuracy of the sample estimates*. Essentially, the loss of accuracy increases with the variance of the weight variables. In case of panel weights, in particular, the variance of the weights can be expected to grow from wave to wave (see equation 1), thus increasingly reducing the accuracy of sample estimates with each additional wave. The situation arises especially in cases, as in TREE, where a large number of successive panel waves are conducted. Frequently, only a few cases with extreme values strongly inflate the variance of weight variables and hence severely impair the accuracy of estimate. According to Kish (1992), the sample variance of a weighted mean (μ_w) compared to an unweighted one (μ) can be described by the following equation where *cv* is the coefficient of variation of the weight variable:

$$\mathrm{var}(\mu_w) = \mathrm{var}(\mu) \cdot \left(1 + cv^2\right) \tag{5}$$

Apart from more or less seriously compromising the accuracy of sample estimates, extremely large individual weights also bear the major disadvantage of having a greatly disproportionate impact on the analyses of smaller subsamples that may render them unstable.

Weight truncation provides a means of avoiding or at least mitigating such adverse effects. This involves trimming individual weights greater than a fixed maximum value to that value. The *optimal threshold value* for truncation is determined by performing an analysis based on equation (5). Table 5 provides a numerical example of such an analysis based on the T4 weights. The weight variable employed in the example is calculated using equation (1) and then recalibrated to a mean of 1.[1] In the first column, the upper threshold value, which the weights are truncated to, is varied systematically for testing purposes. The next column lists the resulting coefficients of variation for the T4 weights truncated at various thresholds. The fourth column gives the increase in the variance of the weighted compared to the unweighted sample estimator, as defined by equation (5), as a function of the truncation threshold values. Without truncation, we would therefore have to expect about a sixfold purely weight-related loss in the accuracy of estimate. The more the weight variability is reduced by way of truncation, the more this unfavourable ratio gradually diminishes. If the *number* of truncated individual weights in the rightmost column is additionally taken into consideration, truncation at a threshold value of 8 proves to be optimal for the example in question. A more extreme truncation results in an only small improvement in the accuracy of sample estimates while at the same time the number of affected weights and respondents climbs sharply, hence compromising the unbiasedness of the sample estimates and the effectiveness of the non-response corrections. There is evidence that the trade-off between accuracy and increase in bias will only render positive results when truncation is cautiously limited to a small number of cases. On grounds of these considerations, the recalibrated raw T4 weights are truncated at an upper

[1] For this purpose, the raw T4 weight is divided by its mean. This recalibration does not affect optimal truncation in any way.

threshold of 8.[1] As Table 5 shows, truncation of only 39 extreme weights (0,8 % of the T4 sample) allows substantially reducing the weight-related loss of accuracy. Owing to truncation, the sampling errors to be expected in accordance with equation (5) are only a good two and a half times instead of six and a half times greater than in the case of an *unweighted* sample of the same size.

Table 5: *Truncation of raw T4 weights*

	cv	var(μ_w)/var(μ)	Number of truncated weights
Without truncation	2.35	6.52	0
Truncation of recalibrated G_i			
> 50	2.12	5.49	2
> 20	1.67	3.79	10
> 10	1.37	2.88	33
> 9	1.32	2.76	36
> 8	1.28	2.64	39
> 7	1.23	2.52	51
> 6	1.18	2.39	67
> 5	1.11	2.24	93

Similarly, truncation can serve to substantially improve the accuracy of weighted sample estimates for all other TREE waves. Based on the criteria exemplified above, eight proves to be the ideal upper threshold value for truncation in case of the weights for the first five waves, seven for the sixth and seventh, and four for the eighth TREE wave.[2] The following table demonstrates how truncation improves the accuracy of sample estimates and documents the number of extreme weights affected.

Table 6: *Truncation, accuracy of estimate, and number of weights affected*

	Without truncation var(μ_w)/var(μ)	With truncation var(μ_w)/var(μ)	Number of affected weights no. cases	(%)
Sample wave 1	2.6	2.1	18	(0.3)
Sample wave 2	2.8	2.2	23	(0.4)
Sample wave 3	8.4	2.4	26	(0.5)
Sample wave 4	6.5	2.6	39	(0.8)
Sample wave 5	7.2	3.2	52	(1.2)
Sample wave 6	55.2	3.5	52	(1.3)
Sample wave 7	118.6	4.9	54	(1.4)
Sample wave 8	199.9	5.2	51	(1.5)

All values are based on the sample realized for each specific wave.

[1] For the raw raising weight as defined by equation (1), this translates into an upper threshold value of about 110.

[2] Said truncation threshold values still refer to the raw weights G_i as defined by equation (1) that have been recalibrated to the mean of 1. Determining the ideal truncation threshold value always involves a considerable degree of discretion; for this reason, the dataset used for weighting also includes raw weights that are not truncated. This enables defining the trade-off between accuracy and bias for each individual analysis.

The results in *Table 6* clearly show that without truncation the accuracy of estimate drops to intolerably low levels from wave 3 and especially from wave 6 on. Except for analyses drawing on data from the first two waves only, the results strongly advise against basing analyses on weights that are not truncated. Even after truncation, weighting still entails a considerable decrease in the accuracy of estimate; accuracy clearly diminishes little by little with each wave just as the increasing variance of weights across waves would lead us to expect. The additional loss in accuracy observed since the seventh wave is particularly remarkable. However, when interpreting the accuracy figures above, we must bear in mind that we are talking about *relative* losses compared with an unweighted sample *of the same size*. The measure of comparison is thus an utmost accurate estimate since the TREE sample is extraordinarily large. Even in the eighth wave the sample still comprises 3,400 respondents – a sample size enabling far more accurate estimates than in many other cases.

5 Poststratification

The weighting factors are constructed, among other things, to enable assessment of the absolute and relative sizes of selected subpopulations in a longitudinal perspective. For descriptive purposes of that kind, a poststratification of those weights is performed to hold the size of a number of particularly important subpopulations constant across all panel waves (see Elliot, 1991; Kish, 1995).[1] Poststratification helps to further stabilize sample estimates.

Table 7: *Reference distribution for poststratification of the sample*

School type lower secondary level	Gender	Language region	Share (%)
Advanced requirements [1)]	female	German	24.9
Advanced requirements [1)]	female	French	9.3
Advanced requirements [1)]	male	German	22.1
Advanced requirements [1)]	male	French	8.8
Basic requirements [2)]	female	German	10.1
Basic requirements [2)]	female	French	2.3
Basic requirements [2)]	male	German	13.9
Basic requirements [2)]	male	French	2.3
Integrated school type	female	Italian [3)]	1.1
Integrated school type	female	German / French	1.8
Integrated school type	male	Italian [3)]	1.4
Integrated school type	male	German / French	2.2
Total			100.0

1) Lower secondary level advanced track (secondary school or *Gymnasium*)
2) Lower secondary level basic track (*Realschule*).
3) Integrated school type is only available option.

[1] To the extent that the weighting model fails to include *all* sources of systematic non-response in a properly specified fashion, the weighted sample distribution may deviate from the initial distribution at the outset.

Since suitable reference distributions, for instance from official statistics, for the school-leaver population of 2000 are missing, poststratification is based on the data collected in wave 1. Those data are least affected by panel mortality and are thus best suited for approximating the unknown distribution for the population in question. Poststratification takes into account school type, gender and language region in defining twelve strata, as shown in the table above, that are held constant in terms of size across panel waves.

Poststratification ensures that the respective recalibrated sample weights for all subsequent waves conform to the reference distribution derived from the sample for wave 1 as displayed in said table.[1]

6 Raising factors and sample weights

Individual panel weights calculated according to equation (1) are not well suited for inferential statistical procedures but only for *estimates* for the population in question (see section 1). A count weighted by factor G_i provides an estimate for the total number of individuals in a population possessing the respective attributes. If one has *inferential statistical analyses* in mind, where significance tests, standard errors and confidence intervals come into play, G_i must be recalibrated in each case so that the sum of the weights corresponds to the size of the sample under analysis (Moser & Kalton, 1971). Statistics software in some cases performs such recalibration automatically; in others, it must be done manually.[2] For this reason, the dataset for each TREE wave also contains recalibrated sample weights in addition to the raising weights. However, we must keep in mind that for inferential statistical analyses based on *subsamples*, or in cases of an appreciable reduction of sample size due to *missing values*, the recalibration must be performed anew.[3] The *dataset containing the TREE weights* consists of four kinds of weight variables for each TREE wave in addition to the design weights for the composite PISA sample described in section 3.1. Apart from a raw raising weight as defined by equation (1), a truncated raising weight poststratified as described in section 5 is also available. All of the truncated and poststratified raising weights are recalibrated to a *population total* of 80,000 in each case. This is an approximation of the underlying population, the actual size of which is not precisely known.[4] For each of the two raising weights, the dataset also contains a corresponding sample weight designed for inferential statistical analyses that differs only in that it is recalibrated to a sample mean of 1. In general, statistical analyses are best performed based on truncated and calibrated sample weights (but also see section 7); the respective raising weights are used only when estimating the *absolute* size of some population is the issue.

Table 8 lists the most important *distribution figures for the thus derived individual weights* required for expansion and statistical inference. The youths participating in a specific wave make up the respective sample (also see Table 1).[5]

---

1 For this purpose, the wave-specific weights are multiplied by a calibration factor specific to each stratum. Two cases where the school type is missing have been excluded from poststratification.

2 This is accomplished by dividing the raising weight by its mean for the respective sample under analysis.

3 The *mean* of the weight variables *for the analyzed sample* should always be 1.

4 Since poststratification and recalibration are performed *after* truncation of weight variables, the calibrated raising and sample weights can take on higher maximum values than the truncation criteria specified in the last section would suggest.

5 All variables listed are included in the data file containing the weight variables (TREE-Weights_T1-T8.sav). In addition, the file includes all predicted probabilities for the wave-specific models as well as the auxiliary variables used in calculating and poststratifying the weights.

Table 8: *Descriptive statistics of raising factors and sample weights**

	Mean	Sum	Standard deviation	Minimum	Maximum	N
Weights for wave 1						
raw raising factor	13.9	76,620	17.8	1.36	691	5,532
raw sample weight	1.0	5,532	1.3	0.10	50	5,532
truncated & calibrated raising factor	14.5	80,000	15.3	1.44	117	5,532
truncated & calibrated sample weight	1.0	5,532	1.1	0.10	8	5,532
Weights for wave 2						
raw raising factor	16.0	83,262	21.4	1.39	463	5,210
raw sample weight	1.0	5,210	1.3	0.09	29	5,210
truncated & calibrated raising factor	15.4	80,000	16.9	1.30	145	5,210
truncated & calibrated sample weight	1.0	5,210	1.1	0.08	9	5,210
Weights for wave 3						
raw raising factor	19.3	94,166	52.4	1.42	2,769	4,882
raw sample weight	1.0	4,884	2.7	0.07	144	4,882
truncated & calibrated raising factor	16.4	80,069	19.4	1.16	147	4,882
truncated & calibrated sample weight	1.0	4,884	1.2	0.07	9	4,882
Weights for wave 4						
raw raising factor	22.5	105,290	52.9	1.45	1,954	4,680
raw sample weight	1.0	4,680	2.4	0.06	87	4,680
truncated & calibrated raising factor	17.1	80,000	22.3	1.14	177	4,680
truncated & calibrated sample weight	1.0	4,680	1.3	0.07	10	4,680
Weights for wave 5						
raw raising factor	27.9	125,779	69.3	1.48	1,527	4,504
raw sample weight	1.0	4,504	2.5	0.05	55	4,504
truncated & calibrated raising factor	17.8	80,000	23.7	0.74	201	4,504
truncated & calibrated sample weight	1.0	4,504	1.3	0.04	11	4,504
Weights for wave 6						
raw raising factor	43.5	180,026	320.4	1.55	15,564	4,135
raw sample weight	1.0	4,135	7.4	0.04	357	4,135
truncated & calibrated raising factor	19.3	80,000	28.5	0.86	262	4,135
truncated & calibrated sample weight	1.0	4,135	1.5	0.04	14	4,135
Weights for wave 7						
raw raising factor	72.7	289,341	788.0	1.62	35,713	3,982
raw sample weight	1.0	3,982	10.8	0.02	491	3,982
truncated & calibrated raising factor	20.1	80,000	35.9	0.47	404	3,982
truncated & calibrated sample weight	1.0	3,982	1.8	0.02	20	3,982
Weights for wave 8						
raw raising factor	163.9	561,171	2,312.1	1.69	97,089	3,424
raw sample weight	1.0	3,424	14.1	.01	592	3,424
truncated & calibrated raising factor	23.4	80,000	47.4	.54	449	3,424
truncated & calibrated sample weight	1.0	3,424	2.0	.02	19	3,424

* The distribution figures listed in the table refer to the wave-specific samples of *participating* youths.

7 Some remarks on the application of the panel weights

In this section, a few issues concerning the application of the TREE panel weights in statistical analysis shall be briefly discussed. The first question to be addressed concerns the appropriate choice of sample weights to match different types of analyses. Subsequently, some issues regarding the *assessment of non-response bias* will be considered as well as questions pertaining to the efficacy of the non-response corrections figured into the sample weights. The brief discussion will then close with a few remarks on the *calculation of sampling errors* and tests of significance based on the weighted panel sample.

Typically, statistical analyses are based on the sample weights of the *most recent* panel wave from which the data to be analyzed originated. For instance, the sample weights of the third wave are employed in analyzing data from PISA and the first three TREE waves. Or in a cross-sectional analysis of a single wave, the sample weights of the respective wave are applied. This is a valid rule of thumb as long as the data for all respondents going into the analysis originate from the same waves. The situation changes if, for instance, the object of research is a transition process that may take place at different points in time for individual respondents and thus is recorded in different panel waves (see e.g. Hupka, Sacchi & Stalder, 2006).

In that case, the probability of sample inclusion and thus the sample weight depends on the specific wave in which the transition is recorded that marks the end of the time period under study in a given respondent's life. For such an analysis, each youth is *individually* assigned the sample weight of the specific wave in which the relevant information on the transition was recorded (or in more general terms: in which the most recent information needed for an analysis was recorded). The *raw* raising factors in the dataset are employed for this purpose since the calibrated weights are not directly comparable across the waves due to the wave-specific calibration constants that are factored in. A raw raising weight newly compiled in this manner must subsequently be truncated and calibrated, as described in section 4 and 6. For the reasons discussed above, truncation should not be neglected in most cases. This is particularly critical if part of the data to be analyzed has been collected in the third wave or later (see end of section 4). Generally speaking, the compilation of individual weights described so far is appropriate when the observation span covers different time periods in the lives of the individual respondents, implying that the most recent data going into the analysis stems from different waves for the individual sample members.

After the appropriate sample weight for a given analysis has been selected (or compiled as described above) and, if necessary, recalibrated (see section 6, FN 1), the question arises as to how corrections for non-response affect estimates. The best way to answer this question is to compare the weighted analyses with an otherwise identical analysis that is based on the – if necessary, recalibrated – design weights for the composite PISA sample (which serves as the starting wave of the TREE panel). If these analyses essentially lead to more or less identical results, this allows the conclusion that the corrections for non-response contained in the panel weights do not substantially influence the estimates. If not, this raises questions as to the plausibility of the observed disparities in light of the findings in section 3.3 (and the full German documentation of the attrition models) and the state of the art in research on non-response and panel attrition. In contrast, the impact of the design weights on the estimates poses no problem in this respect as they exclusively compensate for the complex design of the composite PISA 2000 (see section 1). The differences in results based on design weights and panel weights, however, stem from wave-specific models correcting for attrition bias, which give

rise to both sampling errors and potential specification problems (Menard 2002: 67 f.). Specification problems cannot be ruled out since in constructing the models considerable effort was invested in comprehensively identifying all the attributes of the panel participants and their context that could reasonably be expected to influence panel participation. Such a to some extent inductive approach, on the one hand, does a good job of correcting non-response bias as exhaustively as possible. This minimizes specification errors of the '*omitted variable bias*' type (Menard 2002: 68f.), at least as far as empirically recorded attributes are concerned. On the other hand, this approach inevitably runs the risk of "*model overfit*" – that is, overly adjusting the model to fit purely idiosyncratic sample distributions (also see Wießner 2003: 89). Against this background, it is good advice to determine the impact of the non-response correction on sample estimates as described and to assess its theoretical plausibility.

I would like to close by briefly pointing out that an accurate *estimation of sampling variance* on the basis of the weighted TREE panel requires considering the complex structure of the underlying PISA sample. Even if correctly calibrated panel weights (see section 6) are used, statistical packages that implicitly or explicitly are confined to simple random samples do not allow accurate estimates of sample variance. In general, we would expect them to systematically *underestimate standard errors and confidence intervals* while *overestimating levels of significance* accordingly. Instead, either inductive resampling methods (e.g. bootstrap methods, cf. Mooney & Duval, 1993) or variance estimators designed for complex samples should be used (see Lee, Forthofer & Lorimor, 1989), as implemented in STATA or recent versions of SPSS (STATA: 'svy'-commands; SPSS: 'complex samples'-tools).

References

Elliot, Dave (1991). 'Weighting for Non-Response. A Survey Researcher's Guide'. Office of Population Censuses and Surveys (OCPS), Social Survey Division: London.

Everitt, Brian S. (1993). 'Cluster Analysis'. Edward Arnold: London (3rd edition).

Hagenaars, Jacques A. (1990). 'Categorical Longitudinal Data. Log-Linear Panel, Trend, and Cohort Analysis'. Sage: Newbury Park.

Haisken-DeNew, John & Joachim Frick (2000). 'Desktop Companion to the German Socio-Economic Panel (GSOEP)'. DIW Berlin (4th edition).

Hosmer, David W. & Stanley Lemeshow (1989). 'Applied Logistic Regression'. John Wiley & Sons: New York.

Hupka, Sandra, Stefan Sacchi & Barbara Stalder (2006). 'Herkunft oder Leistung? Analyse des Eintritts in eine zertifizierende nachobligatorische Ausbildung anhand des Jugendlängsschnitts TREE'. TREE: Bern.

Jaccard, James (2001). 'Interaction Effects in Logistic Regression'. 'Sage University Paper Series on Quantitative Applications in the Social Science' Vol. 135, ed. by Michael S. Lewis-Beck. Sage: Beverly Hills.

Kish, Leslie (1992). 'Weighting for Unequal Pi', Journal of Official Statistics, Vol. 8 (2): 183-200.

Kish, Leslie (1995 [1965]). 'Survey Sampling'. John Wiley: New York.

Klein, Sabine & Rolf Porst (2000). 'Mail Surveys. Ein Literaturbericht'. Technischer Bericht 10. ZUMA.

Koch, Achim & Rolf Porst (eds.) (1998). 'Nonresponse in Survey Research'. Proceedings of the Eighth International Workshop on Houshold Survey Nonresponse' ZUMA: Mannheim.

Lee, Eun Sul, Ronald N. Forthofer & Ronald J. Lorimor (1989). 'Analyzing Complex Survey Data'. 'Quantitative Applications in the Social Science' Vol. 71, ed. by Michael Lewis-Beck. Sage: Newbury Park.

Menard, Scott (2002). 'Applied Logistic Regression'.'Sage University Paper Series on Quantitative Applications in the Social Sciences' Vol. 76, ed. by Michael S. Lewis-Beck. Sage: Thousand Oaks (2nd edition).

Meyer, Thomas (2000). 'Evaluation des TREE-Adressblätter-Rücklaufs' TREE: Bern.

Mooney, Christopher Z. & Robert D. Duval (1993). 'Bootstrapping. A Nonparametric Approach to Statistical Inference'. 'Quantitative Applications in the Social Science' Vol. 95, ed. by Michael S. Lewis-Beck. Sage: Newbury Park.

Moser, Claus A. & Graham Kalton (1971). 'Survey Methods in Social Investigation'.Heinemann: London (2nd edition).

PISA Consortium (2000a). 'PISA International Data Base'. OECD (Ed.).

PISA Consortium (2000b). 'PISA Weighting and Variance Estimation'. OECD (ed.): Paris.

PISA, Programme for International Student Assessment (2002). 'Für das Leben gerüstet? Die Grundkompetenzen der Jugendlichen − Nationaler Bericht der Erhebung PISA 2000'. Bundesamt für Statistik, EDK: Neuchâtel.

PISA Romandie (no date). 'La pondération de l'échantillon des élèves de 9e pour l'enquête PISA en suisse romande'.

Renaud, Anne, Erich Ramseier & Claudia Zahner (2000). 'PISA 2000: Sampling in Switzerland. General Information and Design'. PISA.ch (ed.): 'Report for the International Consortium'.

Rizzo, Lou, Graham Kalton & J. Michael Brick (1994). 'Weighting Adjustments for Panel Nonresponse in the SIPP'. Final Report. Westat Inc.: Rockville.

Sacchi, Stefan (2001). 'Longitudinal-Gewichtungen ausgewählter Haushaltspanels. Review im Auftrag des schweizerischen Haushaltspanels'. Cue Sozialforschung: Zurich.

Sacchi, Stefan (2003). 'Longitudinale Stichprobengewichtung fur das TREE-Panel (Befragungswellen 1 & 2)'. Cue Sozialforschung: Zurich.

Sacchi, Stefan (2004a). 'Revision der longitudinalen Stichprobengewichtung für das TREE-Panel (Befragungswellen 1 & 2)'. Cue Sozialforschung: Zurich.

Sacchi, Stefan (2004b). 'Longitudinale Stichprobengewichtung für Welle 3 des TREE-Panels'. Cue Sozialforschung: Zurich.

Sacchi, S. (2008a). Construction of TREE Panel Weights. Documentation for the eight panel waves from 2000 to 2007. Bern/Zurich: TREE & cue sozialforschung.

Sacchi, Stefan (2008b). TREE-Längsschnittgewichtung: Konstruktion und Anwendung. Dokumentation zu den acht Erhebungswellen 2000 bis 2007. TREE and cue sozialforschung: Bern/Zurich.

Sacchi, Stefan (2008c). 'Varianzschätzung mit dem TREE-Panel'. Cue Sozialforschung: Zurich.

Scherpenzeel, Annette (2001). 'Mode Effects in Panel Surveys: A Comparison of CAPI and CATI'. Bundesamt für Statistik (ed.): 'BFS Aktuell': Neuchâtel.

Schnell, Rainer (1997). 'Nonresponse in Bevölkerungsumfragen. Ausmass, Entwicklung und Ursachen'. Leske + Budrich: Opladen.

Stalder, Barbara & Dellenbach, Myriam (2005). 'Dokumentation der Nonresponse-Befragungen im Anschluss an die TREE-Erhebungswellen 2003 und 2004'. TREE-Arbeitspapier: Bern.

Stoop, Ineke (2005). 'The Hunt for the Last Respondent. Nonresponse in Sample Surveys.' (SCP Report Vol. 8). The Hague: Social and Cultural Planning Office.

Wießner, Frank (2003). 'Nonresponse bei Verbleibsuntersuchungen. Korrekturverfahren zu Antwortausfällen am Beispiel ehemals arbeitsloser Existenzgründer, die mit dem Überbrückungsgeld (§57 SGB III) gefördert wurden.' Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 36 (1): 77-96.