ELSEVIER

Studies in Educational Evaluation 31 (2005) 129-144

Studies in Educational Evaluation

www.elsevier.com/stueduc

AN EXPLORATORY ALTERNATIVE APPROACH FOR STUDENT NON RESPONSE WEIGHT ADJUSTMENT[1]

Christian Monseur

University of Liège, Belgium

Abstract

Large scale surveys in education have to face non-response issues that might bias the results. Non-response can occur at three levels: (i) a school refuses to participate, (ii) a sample student fails to participate and (iii) a participating student refuses to answer a particular question. Until now schools and student non-response have been counterbalanced by a non-response weight adjustment. This assumes that both the school and student non-respondents have similar characteristics to the school and student respondents, respectively, within classes. In this article results of analyses conducted on the Student Tracking Form data of the OECD/PISA 2000 survey are presented. The non-randomness of student absenteeism or refusal is demonstrated. Then, a simulation that compares the relative efficiency of the student weight adjustment with a multiple imputation method is presented where the superiority of the multiple imputation method, in particular for educational systems with a small school variance is shown. Finally, the multiple imputation method applied to some PISA 2000 countries identifies biases that become substantial in a longitudinal perspective.

Introduction

This article presents an alternative approach to student weight adjustment for counterbalancing potential biases introduced by student non-responses. It does not deal with school non-response issues. After a short review of the procedure usually used in large-scale national and international assessments, the article will present results of

0191-491X/04/$ - see front matter © 2005 Published by Elsevier Science Ltd.
doi: 10.1016/j.stueduc.2005.05.006

analyses conducted on the *Student Tracking Form* data of the OECD/PISA 2000 survey. It will show the non-randomness of student absenteeism or refusal to participate. Subsequently, results of a simulation comparing student weight adjustment with the proposed alternative approach will be presented. The alternative approach involves drawing cognitive plausible values for absent students. Finally, in the last section it will be shown that, while drawing cognitive plausible values for absent students provides better achievement estimates, it replaces unit student non-response with contextual questionnaire item non-response. Different solutions for overcoming this item non-response will then be discussed.

Current Practice for Student Non Response

Simple random sampling is very rarely used in education surveys because (i) it would be too expensive; (ii) it would not be practical, and (iii) it would be impossible to link, from a statistical point of view, student variables and school, class, or teacher variables (Monseur, in press). Therefore, surveys in education usually draw a student sample in two steps. First, a sample of schools is selected from a complete list of schools containing the student population of interest, and then within the selected schools, a simple random sample of students or classes is drawn.

The choice between a class sample and a simple random sample of students within selected schools mainly depends on the definition of the target population. Large scale international surveys conducted by the International Association for the Evaluation of Student Achievement (IEA) usually define the target population in terms of grade. For instance, the Third International Mathematics and Science Study target population was defined as *all students enrolled in the two adjacent grades that contain the largest proportion of 13-year-olds students at the time of testing* (Martin & Kelly, 1996). With a target population defined in terms of grade, a class sample or a simple random sample of students across classes can be implemented. A within-school random sample of students will provide more precise population estimates but the class sample will allow the analyses of the effectiveness of teacher practices or class characteristics on student achievement.

The OECD Programme for International Student Achievement (PISA) defines the target population in term of age, i.e., *all 15 year-olds student attending educational institutions located within countries* (Adams & Wu, 2002). With such a target population, only a simple random sample of students can be drawn within each selected and participating school, unless all students in the target population are in a single grade.

Unfortunately, large scale surveys have to face non-response issues that might bias the results. In educational surveys, non-response can occur at three levels:

1. A selected school refuses to participate.
2. A sampled student within a participating school fails to complete the instruments for reasons of absence, health, or language problems or simply because he/she refuses to take part in the assessment.
3. A participating student refuses to answer some questions in the background questionnaire, either because he/she does not know the answer, or because he/she refuses to provide the answer.

The first two sources of non-response are usually described as unit non-response, and the third source of non-response as item non-response. Several techniques exist to compensate item non- response: case deletion (pair-wise deletion, list-wise deletion); reweighting; single imputation (unconditional means, unconditional distribution such as the hot deck imputation, conditional means, and conditional distribution); and multiple imputations. (For more information on these different techniques and their respective effectiveness, see Schafer & Graham, 2002.)

In the international databases produced by both the IEA (Gonzalez & Miles, 2001; Gonzalez & Kennedy, 2003) and the OECD (2002), item non-responses in the contextual questionnaires are always mentioned but in no case are they replaced by an imputed value. Statistical analyses reported in international reports usually implement case deletion methods. A few articles discuss alternative approaches to case deletion to deal with item non-response in international databases (see for instance Winglee, Kalton, Rust, & Kasprzyk, 2001).

Up to now, large-scale surveys in education have compensated for unit non response, both at the school and at the student level, by applying a school weight and a student weight adjustment. These weight adjustments share some features of the well-known and often used method denoted as *poststratification* (Bethlelem, 2002). Poststratification consists of comparing the proportion of sampling units in the sample with the proportion of sampling units in the population. Similarly, school and student non-response adjustments consist of comparing the number of sampled units with the number of participating units.

In IEA studies (Martin & Kelly, 1997; Martin, Mullis, & Kennedy, 2003), the school non- response adjustment is equal to:

$$A_{school} = \frac{n_s + n_{r1} + n_{r2} + n_{nr}}{n_s + n_{r1} + n_{r2}}$$

where n_s is the number of originally sampled schools that participated, n_{r1} and n_{r2}, the number of first and second replacement schools, respectively, that participated, and n_{nr} the number of schools that did not participate. Without the distinction between originally sampled schools and both replacement schools, this adjustment factor is simply the ratio between the number of sampled schools and the number of participating schools.

As this adjustment factor is based on the assumption that participating schools are similar to non- participating schools, it is important to group schools according to some criteria and then to apply a school non-response adjustment per group of schools. It is a common practice to group schools according to the explicit and implicit stratification variables used in the school sample frame.

PISA uses a slightly different approach as the numerator estimates the population of 15 year-olds in the group of schools, and the denominator gives the size of the population of 15 year-olds directly represented by participating schools.

The student non-response adjustment follows the same principles. It consists of comparing, per school, or per school and per class if more than one class is sampled within a particular school, the number of sampled students to the number of participating students. In IEA studies, the student non-response is equal to:

$$A_{student} = \frac{S_{rs}^{i,j} + S_{nr}^{i,j}}{S_{rs}^{i,j}}$$

where $S_{rs}^{i,j}$ is the number of eligible students that participated in the jth classroom in the ith schools and $S_{nr}^{i,j}$ is the number of eligible students that did not participate in the jth classroom in the ith schools. As the PISA target population and its within-school sampling procedures slightly differ from IEA studies, the student weight adjustment consequently differs but it follows exactly the same principle. In most cases however this student non-response factor reduces to the ratio of the number of students who should have been assessed to the number who were assessed.

The school and student non-response adjustment factors can therefore be described as an overweighting of some parts of the school and/or student samples that is proportional to the non-response. It supposes that both the school and student non-respondents have similar characteristics to the school and student respondents, respectively, within weighting classes.

The remaining part of this article will closely analyse this assumption at the student level in the case of PISA 2000. Usually, no information or at least no reliable information is available for non- respondent schools. On the other hand, within each participating school, reliable information is collected for each sampled student through the *Student Tracking Form*. In PISA 2000, two variables were included in the *Student Tracking Form*: gender and grade (Adams & Wu, 2002).

A differential participation rate, in itself, is not a concern as long as it has no impact on the survey measures. However, if the differential participation rate variable has any correlation with one of the survey measure, then the survey results will be biased. The size of the bias is proportional to the importance of the differential participation rate and to the correlation between that variable and the survey measures.

Differential Participation Rates in PISA 2000

The PISA 2000 *Student Tracking Form* consists of all sampled students in each participating school. For the purpose of the analyses included in this section, all excluded students and all ineligible students from a PISA perspective were deleted from the files. Therefore, the analyses were performed only on the students who should have participated.

The two student background variables collected in the *Student Tracking Form* were analyzed to identify differential participation rates. As mentioned in the previous section, even if a differential participation rate is observed, it might have no impact on the country performance estimates or on the background information estimates. It mainly depends on the correlation between the variable with the differential participation rate and the performance variables or background variables. Table 1 provides the correlation coefficients between the PISA 2000 Combined Reading Performance and the two variables included in the *Student Tracking Form*.

Table 1: Correlation Between the Combined Reading Performance and the Background Variables Included in the *Student Tracking Form*

	Gender	Grade*
AUS	-0.17	0.24
AUT	-0.14	0.33
BEL	-0.16	0.61
BRA	-0.10	0.51
CAN	-0.17	0.29
CHE	-0.14	0.39
CZE	-0.17	0.23
DEU	-0.15	0.41
DNK	-0.13	0.19
ESP	-0.14	0.55
FIN	-0.28	0.23
FRA	-0.16	0.64
GBR	-0.13	0.04
GRC	-0.19	0.30
HUN	-0.17	0.39
IRL	-0.15	0.24
ISL	-0.20	.
ITA	-0.20	0.35
JPN	-0.18	.
KOR	-0.11	
LIE	-0.15	0.41
LUX	-0.13	0.43
LVA	-0.26	0.36
MEX	-0.12	0.55
NLD	-0.17	0.47
NOR	-0.21	0.09
NZL	-0.21	0.28
POL	-0.18	.
PRT	-0.12	0.72
RUS	-0.21	0.20
SWE	-0.20	0.19
USA	-0.14	0.37

Note: * No correlation coefficients have been provided for ISl, JPN, KOR and POL as only one grade (or close to only one grade) was tested.

All correlation coefficients between Gender and Reading are negative and range from -0.10 to -0.28. A differential participation rate might therefore have an impact on the reading performance estimate. The correlation coefficients between the student grade and his/her performance in reading range from 0.04 in GBR to 0.72 in Portugal. There are six countries with a correlation coefficient higher than 0.50. In some countries, a differential participation rate by grade might therefore have an impact on the reading performance estimate.

Table 2 presents the student participation rates, by gender and by country. The difference in the participation rate by gender ranges from -3.35 in Poland (the participation rate for boys is higher than the participation rate for girls) to 5.16 in Portugal.

Table 2: Participation Rate by Gender

	Girls' participation rate	Boys' participation rate	Difference
AUS	84.77	83.20	1.56
AUT	91.72	91.59	0.13
BEL	94.18	93.25	0.93
BRA	86.94	85.24	1.70
CAN	84.96	82.50	2.46
CHE	95.08	93.94	1.14
CZE	91.83	92.92	-1.09
DEU	84.87	80.39	4.48
DNK	91.74	90.69	1.04
ESP	92.12	91.43	0.69
FIN	93.65	91.92	1.73
FRA	90.16	91.28	-1.13
GBR	79.20	80.53	-1.33
GRC	97.12	96.54	0.58
HUN	95.34	94.26	1.08
IRL	84.28	85.28	-1.01
ISL	88.81	85.42	3.39
ITA	94.11	92.11	2.00
JPN	99.92	100.00	-0.08
KOR	98.73	98.98	-0.25
LIE	96.88	96.36	0.51
LUX	87.61	86.39	1.21
LVA	92.44	88.83	3.61
MEX	94.92	93.00	1.92
NLD	83.77	84.04	-0.27
NOR	89.53	88.85	0.69
NZL	87.48	89.40	-1.93
POL	85.49	88.83	-3.35
PRT	87.75	82.58	5.16
RUS	96.99	95.44	1.55
SWE	89.18	86.80	2.38
USA	84.43	83.38	1.05

Table 3 presents the participation rates by grade and by country.[2] To avoid meaningless participation rates, data are reported for a particular grade if the national sample included at least 100 students in that grade.

Table 3: Participation Rate by Grade

	7	8	9	10	11	12
AUS	.	.	0.80	0.86	0.79	.
AUT	.	.	.	.	.	.
BEL	.	0.89	0.93	0.98	.	.
BRA	0.82	0.89	0.88	0.87	.	.
CAN	.	0.96	0.88	0.86	0.73	.
CHE	.	0.96	0.96	0.86	.	.
CZE	.	0.86	0.91	0.94	.	.
DEU	.	0.78	0.88	0.74	.	.
DNK	.	0.84	0.93	0.83	.	.
ESP	.	0.76	0.84	0.96	.	.
FIN	.	0.91	0.93	.	.	.
FRA	.	0.87	0.91	0.92	0.98	.
GBR	.	.	.	0.84	0.78	0.66
GRC	.	.	.	0.98	0.95	.
HUN	.	.	.	.	.	.
IRL	.	0.73	0.87	0.77	0.87	.
ISL	.	.	.	0.87	.	.
ITA	.	.	0.85	0.95	0.95	.
JPN	.	.	.	0.96	.	.
KOR	.	.	.	0.99	.	.
LIE	.	.	0.98	.	.	.
LUX	.	0.90	0.89	0.83	.	.
LVA	.	.	.	1.00	.	.
MEX	0.86	0.92	0.97	0.95	.	.
NLD	.	.	.	.	.	.
NOR	.	.	.	0.90	.	.
NZL	.	.	.	0.89	0.93	0.93
POL	.	.	0.87	.	.	.
PRT	0.76	0.80	0.87	0.88	.	.
RUS	.	0.91	0.96	0.96	.	.
SWE	.	0.74	0.88	.	.	.
USA	.	.	0.83	0.87	.	.

The most common pattern is an increase in student participation rate as the grade rises. Belgium, Spain and Portugal are typical examples of such a pattern. In Portugal, the correlation between student participation and his/her grade is equal to 0.28. Two countries have exactly the opposite pattern: in Canada and in Great Britain, participation decreases as the grade rises. In some other countries, the participation rate is the highest for the modal grade and lower for the other grades. Australia, Germany and Denmark are in this category.

Alternative Approach to the Student Non-Response Adjustment

The student non-response adjustment can partly reduce any potential bias introduced by the non-randomness of the student non-response. In the previous section, differential participation rates were identified for the gender and grade variables and these were quite important, especially in Portugal. If all schools were single sex schools and if gender was not correlated with any of the other variables that might explain non-participation, then the adjustment would correct the bias on the gender variable and would therefore correct any bias in the survey measures related to the gender differential participation rate. Similarly, if Portuguese schools were only proposing one grade, then the differential participation rate for the grade variable would be corrected. However schools nowadays tend to be coeducational and in most cases, provide instruction for several grades. As the student non-response adjustment is usually implemented at the school level, it is expected that its effectiveness depends on the structure of the educational system, i.e., how similar are students in a particular school?

This section aims to analyse the effectiveness of the student weight adjustment for reducing the bias of student non-response according to the structure of the educational system, and in particular to the intra-class correlation. The efficiency will be analysed in conjunction with the proposed alternative, i.e. conditional multiple imputations. A set of simulations was performed to answer these two questions. Then, the alternative approach was applied to the PISA 2000 countries where a substantial differential participation rate per grade was observed.

In the broadest sense, the student weight adjustment is analogous to a post-stratification where each school is considered as a stratum. Therefore, this adjustment takes into account only the school attended by the student and considers all other things equal. Actually, more information is available for absent students (as we have noted, gender and grade). As shown in a previous section, there are differential participation rates for gender and grade and the student weight adjustment does not use this controlled non-random non-response.

More complex models that integrate all available information on non-respondents can be implemented. With the mixed coefficients multinomial model as described by Adams, Wilson and Wang (1997), plausible values can be imputed for absent students.[3] In the next section, such models will be developed for some PISA 2000 countries. The variables used for the conditioning are: (i) school attended, (ii) gender, and (iii) grade.

Even with more complex models such as these, one cannot assume the identification of a definitive bias with real data as there always remains some uncertainty about the uncontrolled non-randomness of the non-response. The complex model that will be developed later will assume that within a school, within a grade and for each sex, the non-respondent population belongs to the respondent population or, in other words, once non-response is controlled for gender, grade and the school variable, it becomes random. This assumption is certainly questionable, but it is less questionable than the assumption of the student non-response weight adjustment which only takes into account the school variable. Furthermore, the data collected through the *Student Tracking Form* can be extended to reduce the uncontrolled non-randomness of the non-response. With a class sample, for instance, it would be appropriate to collect the teacher marks in the domain assessed by the international survey. This limitation entirely justifies the use of simulations. Indeed, with the simulation methodology, there is no uncertainty about the population parameters. Therefore, any difference between the estimates and the population parameters can be reported as bias.

The simulation requires modelling student non-response. In this set of simulations, the propensity to participate is simply a function of student proficiency. For each simulation, a population of 15 strata was generated with 200 schools per stratum and with 150 students per school. One parameter was used to generate this population of 450 000 students: the percentage of variance between schools. The mean and the standard deviation of the performance distribution were set at, respectively, 500 and 100. For each simulation, 500 samples were drawn and the results of each sample were combined. For a particular sample, 10 schools from each stratum were selected according to a simple random procedure. Within a selected school, 35 students were randomly selected. According to their readiness to participate, about 20% of the students were excluded.

For each sample, the following statistics for three different methods were computed: (i) the mean and standard deviation before student non-response weight adjustment; (ii) the mean and the standard deviation after the student non-response weight adjustment; (iii) the mean and the standard deviation if plausible values were drawn for absent students, and (iv) the decomposition of the total variance into the school variance and within school variance for the same three methods.

Three parameters were taken into account for this set of simulations:

1. The percentage of variance between schools; four values were used: 0%, 20%, 40% and 60%.
2. The correlation between student proficiency and the student's readiness to participate. Two values were used : 0.2 and 0.4.
3. The correlation between the student information provided on the student tracking form and reading proficiency. Ten values were used: 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0

As the PISA 2000 international standard for student participation was set at 80%, we did not vary the overall student response rate from one simulation to another. All the data presented in this section are based on an expected overall student response rate of 80%.

Table 4: Means and SE: Rho of 0.0 and Correlation of 0.20 Between Proficiency and Participation

	No adjustment		*Weight adjustment*		*PV adjustment*	
Corr	Mean	SE	Mean	SE	Mean	SE
0.9	507.04	(1.484)	507.05	(1.493)	501.42	(1.402)
0.8	507.05	(1.513)	507.05	(1.516)	502.67	(1.420)
0.7	507.02	(1.508)	507.02	(1.520)	503.73	(1.514)
0.6	506.97	(1.567)	506.97	(1.561)	504.52	(1.570)
0.5	506.95	(1.611)	506.95	(1.615)	505.27	(1.645)
0.4	507.08	(1.541)	507.08	(1.545)	506.03	(1.621)
0.3	507.02	(1.545)	507.01	(1.560)	506.44	(1.691)
0.2	506.99	(1.507)	506.99	(1.516)	506.74	(1.626)
0.1	506.96	(1.458)	506.97	(1.463)	506.93	(1.544)
0.0	506.92	(1.488)	506.92	(1.496)	506.92	(1.604)

The weight adjustment for student non-response is equal to the ratio between the sum of the weight of the selected students within a particular school divided by the sum of the weight of the participating students within that particular school. In the simulation presented in this section, the weight adjustment is therefore equal to 35 divided by the number of participating students. Tables 4-7 present, for each simulation, the mean before adjustment, the mean after adjustment for student non-response, and the mean if plausible values were generated for absent students.

Table 5: Means and SE: Rho of 0.2 and Correlation of 0.20 Between Proficiency and Participation

	No adjustment		*Weight adjustment*		*PV adjustment*	
Corr	Mean	SE	Mean	SE	Mean	SE
0.9	506.79	(3.795)	505.42	(3.781)	501.13	(3.768)
0.8	507.13	(4.047)	505.75	(4.055)	502.54	(4.026)
0.7	507.30	(3.910)	505.92	(3.907)	503.54	(3.972)
0.6	507.03	(3.685)	505.64	(3.684)	503.92	(3.756)
0.5	507.17	(3.937)	505.80	(3.950)	504.66	(4.006)
0.4	507.31	(3.666)	505.94	(3.670)	505.23	(3.673)
0.3	506.87	(4.001)	505.53	(3.983)	505.12	(4.035)
0.2	506.70	(3.665)	505.32	(3.693)	505.15	(3.743)
0.1	507.10	(3.525)	505.72	(3.533)	505.67	(3.593)
0.0	506.97	(3.817)	505.61	(3.824)	505.58	(3.891)

Table 6: Means and SE: Rho of 0.4 and Correlation of 0.20 Between Proficiency and Participation

	No adjustment		*Weight adjustment*		*PV adjustment*	
Corr	Mean	SE	Mean	SE	Mean	SE
0.9	506.76	(5.061)	504.03	(5.082)	501.06	(5.139)
0.8	506.71	(5.095)	503.99	(5.083)	501.80	(5.069)
0.7	507.11	(5.243)	504.30	(5.240)	502.78	(5.289)
0.6	507.24	(5.051)	504.45	(5.069)	503.38	(5.101)
0.5	506.78	(4.922)	504.04	(4.892)	503.38	(4.922)
0.4	506.62	(5.022)	503.86	(5.043)	503.45	(5.085)
0.3	506.89	(5.201)	504.14	(5.202)	503.92	(5.213)
0.2	507.08	(4.676)	504.33	(4.711)	504.24	(4.740)
0.1	506.89	(5.164)	504.14	(5.170)	504.13	(5.189)
0.0	507.08	(4.731)	504.28	(4.766)	504.25	(4.807)

Table 7: Means and SE: Rho of 0.6 and Correlation of 0.20 Between Proficiency and Participation

	No adjustment		*Weight adjustment*		*PV adjustment*	
Corr	Mean	SE	Mean	SE	Mean	SE
0.9	506.72	(6.524)	502.58	(6.592)	500.81	(6.619)
0.8	506.87	(6.567)	502.73	(6.599)	501.61	(6.617)
0.7	506.87	(6.414)	502.74	(6.350)	501.97	(6.390)
0.6	507.14	(6.296)	502.98	(6.331)	502.48	(6.328)
0.5	506.69	(6.262)	502.56	(6.248)	502.25	(6.256)
0.4	506.71	(6.484)	502.58	(6.440)	502.40	(6.440)
0.3	506.78	(6.243)	502.65	(6.256)	502.55	(6.234)
0.2	507.33	(6.051)	503.20	(6.078)	503.18	(6.122)
0.1	507.00	(6.411)	502.88	(6.422)	502.87	(6.410)
0.0	506.55	(6.006)	502.43	(6.086)	502.41	(6.096)

From these tables and other results presented in Monseur and Wu (2002), several conclusions can be drawn:

1. The efficiency of weight adjustment for student non-response is proportional to the between-school variance. If the school variance is equal or close to 0, then the weight adjustment does not contribute to reducing the bias. On the other hand, the results of the simulations show that the bias for student non-response can be reduced by more than 50% if the percentage of variance between schools is large.
2. Whatever the intra-class correlation and whatever the correlation between student readiness to participate and his/her proficiency, the plausible value method appears, depending on the correlation between the *Student Tracking Form* information and the student proficiency, to be equal to or more effective than the weight adjustment for student non-response. The plausible values method is particularly useful when the school variance is small, but the advantage in comparison with the weight adjustment becomes smaller as the school variance increases.
3. The plausible value method is most effective when there is a high correlation between the *Student Tracking Form* information and student proficiency. Table 1 shows that the grade information in highly tracked systems might correlate with achievement at 0.5 or even above for countries like Belgium, Brazil, Spain, France, Mexico and Portugal. On the other hand, in North European countries, the variables collected in the *Student Tracking Form* are not highly correlated with achievement.

Efficiency of the Alternative Approach on the PISA 2000 Data

In PISA 2000, all the variables collected through the international compulsory student questionnaire were included in the imputation of plausible values as conditioning variables. The PISA 2000 conditioning variables were prepared using procedures based on those used in the United States National Assessment of Educational Progress (Beaton, 1987) and in TIMSS (Macaskill, Adams, & Wu, 1998). For more information, see Adams and Wu, 2002.

A new set of plausible values has been drawn for a few PISA 2000 countries where a substantial differential participation rate was observed for gender and/or for grade: BRA, DEU, ESP, GBR, IRL, ITA, LUX, PRT and USA. This new set of plausible values was imputed based on all eligible and not excluded 15 year-old students with the following set of conditioning variables: (i) school attendance, (ii) gender and (iii) grade.[4] Table 8 presents the mean and standard deviation of the combined reading scales computed on the PISA 2000 plausible values and the new set of plausible values. Imputation errors and standard errors are provided for the PISA 2000 estimates.

Table 8: Means, Standard Deviations for Different Models

Country	*Statistics*	*Model 1* *PISA 2000*			*Model 2* *Sampled students*
		Estimates	Imputation Error	Standard Error	
BRA	N	4893			5627
	Mean	396.02	(0.49)	(3.1)	394.48
	STD	86.18	(0.54)	(1.9)	89.00
DEU	N	5073			6023
	Mean	483.99	(0.52)	(2.5)	484.27
	STD	111.21	(0.74)	(1.9)	111.11
ESP	N	6214			6764
	Mean	492.55	(0.40)	(2.7)	489.20
	STD	84.96	(0.47)	(1.2)	86.33
GBR	N	9340			11 562
	Mean	523.44	(0.29)	(2.6)	522.26
	STD	100.49	(0.15)	(1.5)	100.76
IRL	N	3854			4555
	Mean	526.67	(0.46)	(3.2)	524.86
	STD	93.57	(0.43)	(1.7)	94.28
ITA	N	4984			5369
	Mean	487.47	(0.48)	(2.9)	487.02
	STD	91.41	(0.41)	(2.7)	90.71
LUX	N	3528			4060
	Mean	441.25	(0.62)	(1.6)	446.21
	STD	100.44	(0.67)	(1.5)	105.27
PRT	N	4585			5381
	Mean	470.15	(0.58)	(4.5)	462.73
	STD	97.15	(0.33)	(1.8)	100.71
USA	N	3846			4568
	Mean	504.42	(0.49)	(7.1)	498.13
	STD	104.78	(0.54)	(2.7)	106.55

The differences between the two mean estimates show biases that range between -4.96 in Luxemburg to 7.42 in Portugal. If we apply the Graham and Schafer rule, in Spain, Ireland, Luxemburg, Portugal and in the United States the difference is higher than half a standard error. Furthermore, in five countries, the difference between the two standard deviation estimates is higher than 0.5 standard errors. The bias is particularly high in Brazil, Luxembourg, Portugal and the United States.

Applying the student non-response adjustment or drawing plausible values for absent students makes a significant difference from a statistical point of view.

Disadvantages of the Alternative Approach

With this alternative approach, it is necessary to analyse performance data with the student weight before adjustment. However, absent students included in the analyses of performance data will have missing values on all student contextual questionnaire variables. As the student adjustment might reduce some bias, using the student weight before adjustment on context questionnaire data will suppress any chance of reducing bias due to student non-response.

Table 9 provides, for some of the continuous variables in the PISA 2000 database, the standardized difference[5] between two mean estimates. Standardized differences higher than 0.5 or lower than -0.5 are reported in italics.

Table 9: Bias on Continuous Context Variables

	BRA	DEU	ESP	GBR	IRL	ISL	ITA	LUX	PRT	USA
Achievement Press	*0.95*	-0.22	0.19	-0.23	0.11	0.04	0.03	0.34	-0.27	0.06
Sense of belonging	0.04	0.40	0.08	0.16	0.13	-0.01	0.07	*0.64*	-0.21	0.07
Occupation expectations	0.30	*1.66*	0.17	*0.56*	0.28	-0.06	*0.56*	-0.34	-0.12	-0.13
Family Educa. support	-0.39	-0.17	0.06	-0.30	-0.20	0.05	-0.08	-0.10	-0.13	0.02
Home Educa. resources	0.45	*0.84*	0.21	0.34	0.37	0.09	0.26	*1.09*	-0.09	0.10
HISEI	*0.76*	*1.43*	0.26	*0.53*	0.37	-0.26	0.49	0.22	-0.28	-0.03
Homework time	*1.33*	*0.79*	0.17	*0.52*	0.15	0.24	0.34	0.11	-0.22	-0.04
Reading engagement	0.40	*1.13*	0.08	0.16	-0.04	0.02	0.22	0.14	-0.19	-0.11
Social communication	-0.03	0.49	0.05	0.38	0.19	-0.01	0.14	0.14	-0.30	0.18
Wealth	*0.76*	*0.61*	0.22	0.26	0.24	0.05	0.31	*0.60*	-0.35	0.14

As shown in Table 9, using the student weight before adjustment or using the weight after adjustment makes differences in some countries like Brazil, Germany, the United Kingdom and Luxembourg. In the other countries, it does not make any difference. This does not mean that context questionnaire data estimates are unbiased; it does mean that the student weight adjustment is unable to correct a potential bias. It also shows that the student weight before adjustment cannot be used for analyzing the context questionnaire data. To overcome this conflict, there are two possibilities: (i) providing two weights in the database or (ii) imputing the context questionnaire data. Imputing questionnaire data within the international survey time constraints might be a challenge. Further, it would have some consequences on how questionnaire data need to be analyzed. As Winglee et al state (2001), "(...) when analysing an imputed data set, it needs to be recognized that the standard errors of the estimates are larger than those that would apply if there were no missing data". As suggested by Rubin (1987), multiple imputations should be used to obtain unbiased estimates of standard errors. However, international databases are currently complex enough to deter one from adding more difficulty. The other alternative would be to provide the student weight before adjustment and the student weight after adjustment. In the PISA context, this would mean that the international database would need to include two sets of 80 weight replicates as each replicate in PISA 2000 involves its own non-response adjustment (both school and student non-response adjustments). This could be achieved through two data files: the first one would consist of the whole national sample with the cognitive plausible values, the student weight before adjustment and the replicates derived from this weight; the second one would include the subset of participant students with the cognitive plausible values, the questionnaire data and derived variables, the student weight after adjustment and the replicates derived from this weight. As IEA studies are using the Jackknife replication method for stratified samples without specific non-response adjustment for each replicate, this issue can be resolved by providing two weights: the student weight before adjustment (for the analysis of performance data) and the student adjusted weight (for the analysis of questionnaire data, included their relationship with performance data). Whatever the replication method, this solution does not simplify the use of international databases for secondary analyses.

Conclusions

Through a set of simulations and through the analyses of the PISA 2000 *Student Tracking Form*, it has been shown that drawing cognitive plausible values for absent students provides more reliable performance population estimates than the weight adjustment method. However, this alternative approach raises two important issues: (i) the analyses of background variables cannot be solved without substantially increasing the complexity of secondary analyses, and (ii) a bias of about 0.10 standard deviation will not change the major outcomes of the survey and can thus be considered as negligible from a policy point of view. In this context, one might argue that this explanatory alternative should be set aside. Yet this would not reflect the impact of such survey outcomes on the public and on policy makers, in particular with regard to trend indicators. First of all, one should not underestimate the *horse race* character of such a survey. Secondly, while a bias of 0.10 standard deviation can be considered as negligible for a one shot assessment it becomes substantial in a longitudinal perspective. International surveys such as PISA, TIMSS and PIRLS are nowadays designed to provide achievement trend indicators from one data collection to the next. From previous surveys, and especially from TIMSS 99, (Mullis et al., 2000, Martin et al., 2000), changes in the country performance means during a four year period ranged from -26 to 17 with an average of 2 in mathematics, and from -27 to 27 with an average of 3 in science. The standard errors for these trends indicators respectively ranged from 2.2 to 9.5 in mathematics and from 3.3 to 9.8 in science . A bias of 10 points might be interpreted as a significant shift between two data collections.

From our point of view, the PISA 2003 OECD initial report (2004) and the TIMSS 2003 report lack caution when reporting trends indicators. There is no doubt that scholars, educators, policy makers and journalists will generally overstate such results. The consequences of such overstatements will become apparent only after a few years. The policy importance of trend indicators is incompatible with uncertainties such as the bias identified in this article.

Notes

1. The author would like to thank Keith Rust, Juliette Mendelovits and Dominique Lafontaine for their comments.
2. Results for Austria, Hungary and the Netherlands are missing as the grade information was only provided for participating students. In Belgium, the Flemish community did not provide grade information for non-participating students. The student participation rate by linguistic community differs: 90% for the French community, 95% for the Flemish community. The participation rates by grade for the French part of Belgium are respectively 86%, 88% and 95%.
3. As no cognitive information is available for these students, the plausible values will be drawn from the *a posteriori* distribution specific to the cross tabulation of the variables included in the conditioning.
4. The grade information was recoded as dummy variables to avoid forcing the linearity of achievement across grades.
5. Mean with base weight (i.e. the weight before adjustment) minus mean with final weight (i.e. the weight after the adjustment), divided by the standard error.

References

Adams, R.J., Wilson, M.R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-24

Adams, R.J., & Wu. M. (Eds.) (2002). *PISA 2000 Technical Report.* Paris: OECD.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-1984 technical report.* (Report No 15-TR-20). Princeton, NJ: Educational Testing Service.

Bethlelem, J.G. (2002). Weighting non response adjustments based on auxiliary information. In R.M. Groves, D.A. Dilman, J.L. Eltinge, & R.J.A. Little (Eds.), (2002). *Survey nonresponse.* New York: Wiley.

Ganzemboom, H.B.G., de Graaf, P.M., & Treiman, D.J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1-56.

Gonzalez, E.J., & Kennedy, A.M. (2003). *PIRLS 2001 User guide for the international database.* Chestnut Hill, MA: Boston College.

Gonzalez, E.J., & Miles, J.A. (2001). *TIMSS 1999 User guide for the international database.* Chestnut Hill, MA: Boston College

Martin, M.O., & Kelly, D.L. (1996). *Third International Mathematics and Science Study: Technical report, Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

Macaskill. G., Adams, R.J., & Wu, M.L. (1998). Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scale. In M. Martin & D.L. Kelly (Eds.), *Third International Mathematics and Science Study, technical report Volume 3: Implementation and analysis.* Chestnut Hill, MA: Boston College.

Martin, M.O., & Kelly, D.L. (1997). *Third International Mathematics and Science Study: Technical report, Volume II: Implementation and analysis.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., & Kennedy, A.M. (2003). *PIRLS 2001 technical report.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 international science report.* Chestnut Hill, MA: Boston College.

Monseur, C., & Wu, M. (2002). *Imputation for student nonresponse in educational survey.* International Conference for Improving Surveys, Copenhagen, August 25-27 2002.

Monseur, C. (in press). *PISA database user guide.* Paris: OECD.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J.., & Smith, T.A., (2000). *TIMSS 1999 international mathematics report.* Chestnut Hill, MA: Boston College

OECD (2002). *Manual for the PISA 2000 database.* Paris: OECD.

OECD (2004).PISA 2003 initial report. Paris: OECD.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rust, K. (2000). *Pisa meeting and variance estimation.* Document presented at a PISA 2000 meeting in Rome.

Schafer, J.L., Graham, J.W. (in press). Missing data: Our view of the state of the art. *Psychological methods.*

Winglee, M, Kalton, G., Rust, K., & Kasprzyk, D. (2001). Handling item nonresponse in the U.S. Component of the IEA Reading Literacy Study. *Journal of Educational and Behavioral Statistics, 26,* 343-359.

The Author

CHRISTIAN MONSEUR is a researcher at the Institute of Educational Sciences at Liège University in Belgium. He is also an associate researcher for the Australian Council for Educational Research where he works in areas of sampling and weighting for the OECD/PISA Study. Before assuming these roles, Christian was the data manager for PISA 2000 and director of the PISA Plus project. He has a qualification as a teacher, has graduated in Educational Sciences and has completed a Masters degree in Statistics. He has published a number of articles and chapters in the field of educational assessment.

Correspondence: <cmonseur@ulg.ac.be>