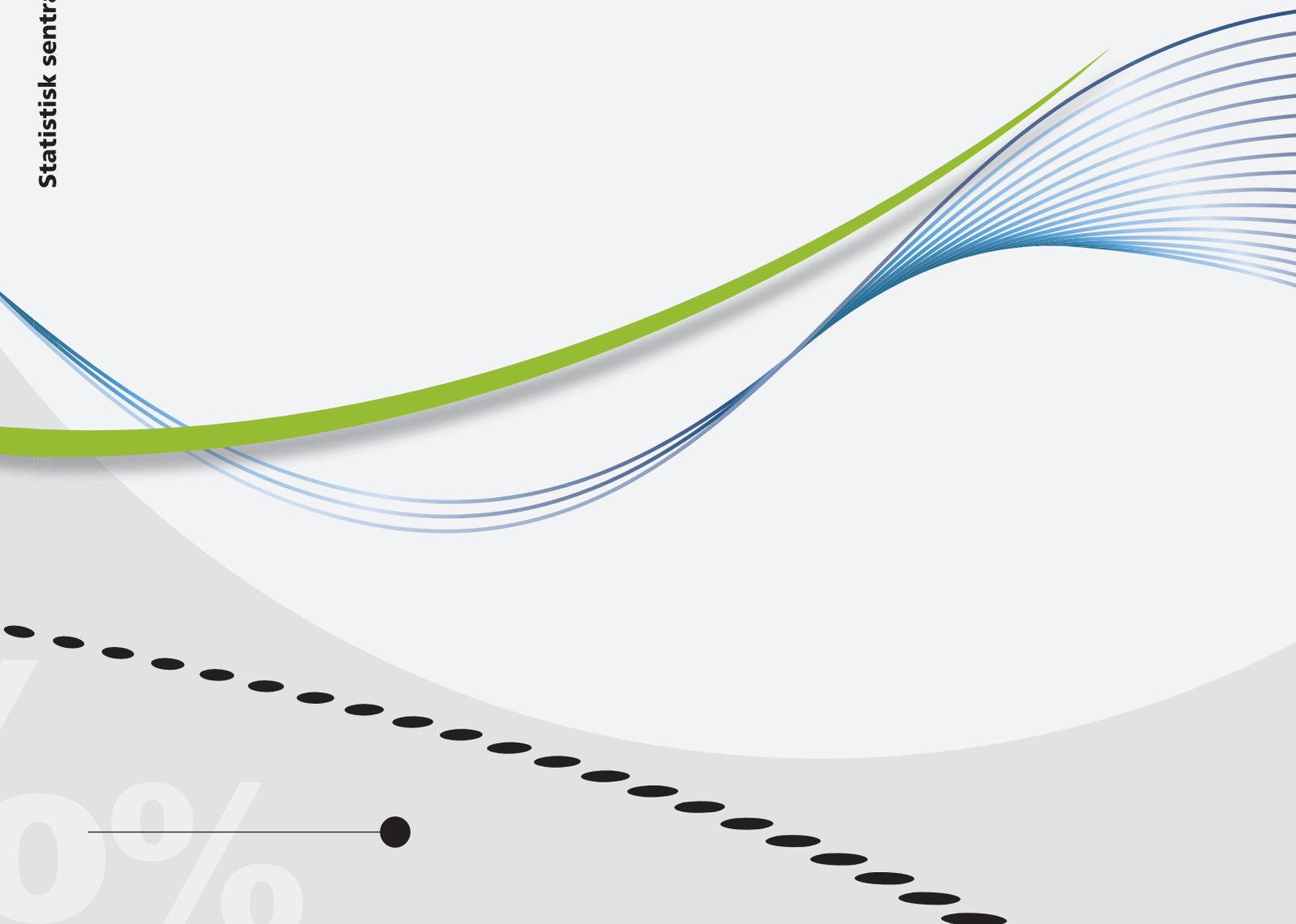


Hilde Eirin Pedersen and Elisabeth Falnes-Dalheim

**Non-response and representativity
in a survey on education completed
abroad**

Q2012 Conference in Athens in May 2012



Hilde Eirin Pedersen and Elisabeth Falnes-Dalheim

**Non-response and representativity in a survey
on education completed abroad**

Q2012 Conference in Athens in May 2012

© Statistics Norway When using material from this publication, Statistics Norway shall be quoted as the source.	Symbols in tables	Symbol
	Category not applicable	.
	Data not available	..
	Data not yet available	...
	Not for publication	∴
	Nil	-
ISBN 978-82-537-8403-8 Printed version	Less than 0.5 of unit employed	0
ISBN 978-82-537-8404-5 Electronic version	Less than 0.05 of unit employed	0.0
ISSN 1891-5906	Provisional or preliminary figure	*
Subject: 04.90	Break in the homogeneity of a vertical series	—
Published June 2012	Break in the homogeneity of a horizontal series	
Print: Statistics Norway	Decimal punctuation mark	.

Preface

The authors of this article are Hilde Eirin Pedersen, senior executive officer, and Elisabeth Falnes-Dalheim, adviser, in Division for survey planning and user testing, Statistics Norway. The paper was presented at the European Conference on Quality (Q2012) 30 May, 2012, Athens, Greece by Elisabeth Falnes-Dalheim

Statistics Norway, 11 June 2012.

Hans Henrik Scheel

Abstract

The population of this mandatory Survey on education is immigrants with unknown education; the missing group in the Norwegian education register. In the 2011 survey on education completed abroad, more than 200 000 immigrants had to be contacted in order to fill the gap in the register. We know from other surveys by Statistics Norway that immigrants in general have a higher non-response rate than the rest of the population, especially those with unknown education. We expect the non-response rate to vary within this group. We want to compute the R-indicator and see whether the representativity is increasing or decreasing with regard to auxiliary variables such as country background, age, gender and years living in Norway, after the reminders. Lastly, we will discuss the effects of the reminders on the estimation of the level of education in order to conclude whether the results from the survey can be used for statistical purposes without further adjustments.

Contents

Preface	3
Abstract	4
1. Introduction	6
2. Theoretical framework	6
2.1. Non-response and representativity	6
2.2. R-indicator	7
2.3. Unconditional partial R-indicator	7
3. Results	8
3.1. The R-indicator	8
3.2. The partial R-indicators.....	9
3.3. The level of education.....	12
4. Discussion	14
4.1. Did the reminders improve representativity?	14
4.2. Did the reminders affect what responses were received in terms of education level? .	14
5. Conclusion	15
References	16
List of figures	17
List of tables	18

1. Introduction

An important issue when conducting a survey is the response rate. The result of the data collection is often considered better the higher the final response rate that is achieved. However, a higher response rate does not necessarily mean higher representativity.

In this paper we examine closely the survey on education completed abroad conducted in 2011. The main objective is to analyse the representativity by means of the R-indicator and to establish whether sending reminders improved or diminished the representativity with respect to auxiliary variables such as age, gender, occupational status, country background and years living in Norway. We will also discuss how the reminders affect what responses are received in terms of education level.

About the survey

More than 200 000 immigrants living in Norway were sent a letter from Statistics Norway in autumn 2011 in order to fill the gap in the Norwegian education register. The survey was mandatory and went out to every immigrant aged 18 years or older with missing information in the education register. Controlling for registered deaths and remigration, the gross sample was 213 756 persons. Those who did not answer the first letter were sent a reminder. Those who still did not answer received a second reminder. Sixty-one per cent answered either the online survey or the postal survey.

We know from other surveys by Statistics Norway that immigrants in general have a higher non-response rate than the rest of the population, especially those with unknown education. However, immigrants are not a homogeneous group, and we expect the non-response rate to vary depending on different characteristics. We also know that persons with a high level of education more frequently respond to our surveys. We therefore expect there to be an overrepresentation of persons with a high level of education among the respondents.

Table 1.1. Key figures from the survey

Key figures from the education survey	(N)	%
Population	218 336	
Deaths and emigration (per 5 October 2011)	4 570	2
Gross population	213 756	
Postal returns	21 759	10
Non-respondents (postal returns excluded)	61 179	29
Non-respondents (postal returns included)	82 938	39
Net sample	130 818	61
Contact sample (gross sample - postal returns)	191 997	90

2. Theoretical framework

2.1. Non-response and representativity

We divide the gross sample into two main groups: the respondents and the non-respondents. The non-respondents include those who did not answer the survey and those who failed to specify their level of education. The non-response rate is calculated from the gross population and is an indicator of the quality of the survey. However, the non-response rate does not tell us much about the representativity.

Scouten et al. (2009) defines the response subset as representative if the individual response probabilities are equal for all units in the population and if one unit's response probability is independent of all other units' response probability. Because the individual response probabilities are unknown, Scouten et al. (2009) uses a weaker definition: "A response subset is representative of a categorical

variable X with H categories if the average response propensity over the categories is constant.”

2.2. R-indicator

Schouten and Cobben (2007) suggest an indicator for representativity; the R-indicator. The formula used is given below:

$$R(X) = 1 - 2S(\rho_X) = 1 - 2\sqrt{\frac{1}{N-1} \sum_i^N (\rho_i - \bar{\rho})^2}$$

R(X) is between 0 and 1, where 0 is the maximum deviation from representativity and 1 is the strongest representativity.

N is the population size, U the population units, i the label of the population unit, ρ_X is the individual response propensities and $S(\rho_X)$ the standard deviation of the individual response propensities.

The RISQ Project¹ suggests an SAS code to compute the R-indicators. The response probabilities are estimated by logistic regression (Schouten and Shlomo, 2010). An adjusted R-indicator is computed in addition to the unadjusted R-indicator.²

2.3. Unconditional partial R-indicator

Schouten and Skinner (2010) present the unconditional partial R-indicators at both the variable level and the category level. The formulas used are given below.

The category level indicator

The partial R-indicators for category k:

$$P_u(Z, k) = \sqrt{\frac{N_k}{N} (\bar{p}_{Z_k} - \bar{p})}$$

N_k is the number of population units in category k. Z represents the categorical variable.

The category level indicator is between -1 and 1. Negative values indicate underrepresentation and positive values indicate overrepresentation.

The variable level indicator

Schouten and Skinner (2010) also present the unconditional partial indicator for variable X, which measures the variation between the response categories of the k categories:

$$P_u(Z) = \sqrt{\frac{1}{N} \sum_i^K P_u^2(Z, k)}$$

As for the category level indicator, the variable level indicator is between -1 and 1. Negative values indicate underrepresentation of category k while positive values indicate overrepresentation.

¹ The RISQ Project is financed by the 7th Framework Programme (FP7) of the European Union and gives a thorough presentation of the R-indicator. <http://www.risq-project.eu/>.

² We have computed the R-indicators using both the formulas (with the actual response rates) and the SAS code. The results are the same for the unadjusted R-indicators. In this survey, the adjusted R-indicators were equivalent to the unadjusted R-indicators at the level of accuracy required in this analysis.

3. Results

3.1. The R-indicator

In the model below we have included seven Xs, or auxiliary variables: gender, age group, continent, years living in Norway, reason for immigration, household status and work status.

Table 3.1 shows the overall R-indicator after each of the three dispatches.

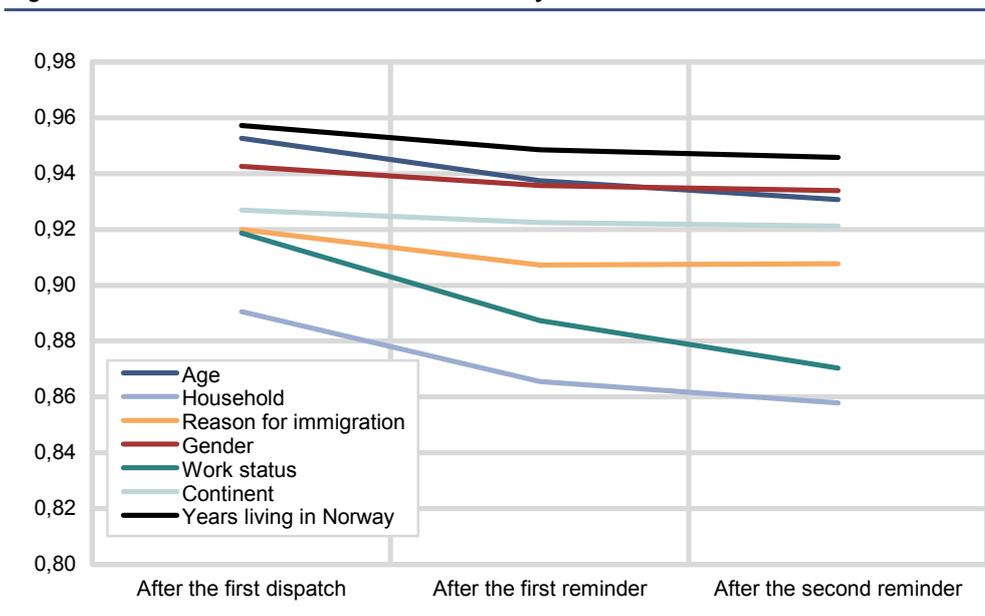
Table 3.1. The overall R-indicator after each dispatch

	R-indicator adjusted	R-indicator unadjusted	Propmean	SE_r	LB_r	UB_r
After the first dispatch	0.82	0.82	0.350	0.002	0.816	0.823
After the first reminder	0.78	0.78	0.511	0.002	0.779	0.787
After the second reminder	0.77	0.77	0.612	0.002	0.770	0.778

The first column shows the R-indicator adjusted for bias. The second column shows the unadjusted R-indicator. Propmean is the response rate. SE_r is the estimated standard error of the R-indicator. LB_r and UB_r are the lower and upper bound of the 95 per cent confidence interval of the R-indicator based on a normal approximation. The R-indicator decreases after each dispatch, indicating a weaker representativity, although the decrease is relatively small. The greatest decrease is found between the first dispatch and the first reminder. Because of the size of this survey (213 756, the gross population), the variance of the R-indicator is very small and the confidence intervals after each dispatch do not overlap. The R-indicators are not independent of each other. We cannot, therefore, conclude that the decrease is significant, but a clear pattern can be seen.

Figure 3.1 shows the R-indicator for each of the seven variables during the data collection process. The same pattern is found here as for the overall R-indicator. The greatest decline in the R-indicator is between the first dispatch and the first reminder. Except for the variable Reason for immigration, there is a decline for every variable after each dispatch, although the decline is relatively small. The variable Work status shows the greatest decline. The variable Household has the lowest R-indicator, i.e. the weakest representativity, through the data collection process, while Years living in Norway has the strongest representativity.

Figure 3.1. The R-indicators for the seven auxiliary variables



3.2. The partial R-indicators

The partial R-indicators include the variable level indicator and the category level indicator.

From table 3.2 we see that all the variable level indicators increased after the first reminder. After the second reminder, most of them increased further and some remained at the same level. This increase indicates that the representativity of the response worsened as the response rate increased.

Table 3.2. Partial R-indicators: the variable level indicators (in bold) and the category level indicators for each of the seven auxiliary variables. Multiplied by 1 000

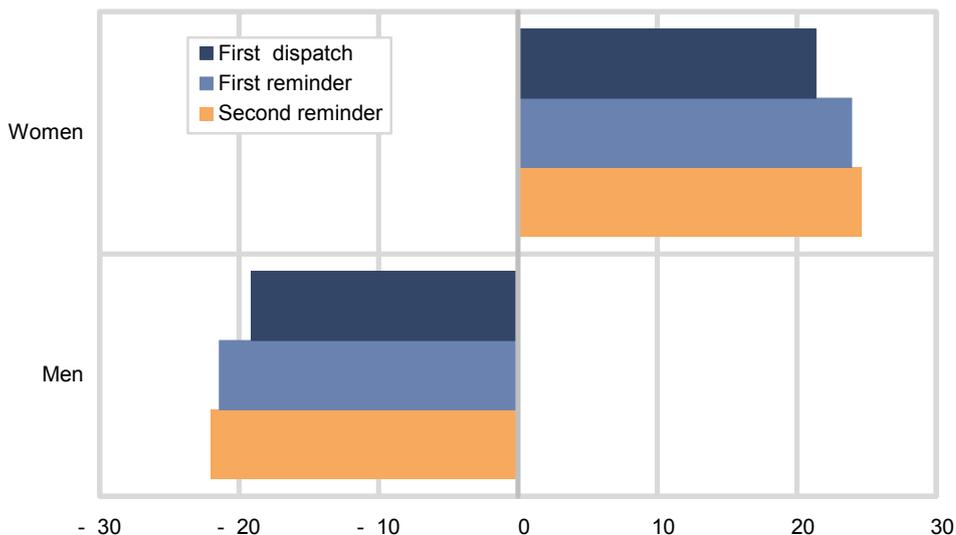
	After the first dispatch	After the first reminder	After the second reminder
	*1 000		
Age	24	31	35
Under 30	-8	-11	-11
30-44	9	10	10
45-66	10	14	16
67 and older	-18	-24	-27
Household	55	67	71
Single without children	-30	-39	-44
Single with children	-15	-19	-17
Couple without children	41	46	44
Couple with children	14	23	30
Reason for immigration	40	46	46
Missing	-17	-23	-25
Other	7	7	6
Work	14	15	12
Family	9	15	20
Refugee	-29	-32	-30
Education	12	8	4
Unknown	-2	-1	2
Gender	29	32	33
Male	-19	-21	-22
Female	21	24	25
Work status	41	56	65
Missing	-22	-34	-41
Not working	-21	-26	-27
Working	27	37	42
Continent	37	39	39
Africa	-23	-27	-32
Asia	-22	-20	-15
Europe	18	19	16
Latin America	3	5	2
Oceania	-1	-1	-1
USA, Canada, Australia, NZ	3	4	5
Unknown	-2	-3	-2
Years living in Norway	21	26	27
Under 2 yrs	-7	-13	-17
2-4 yrs	17	21	20
5 yrs or more	-10	-9	-3

In the following we will take a closer look at the partial R-indicators for each of the seven auxiliary variables.

Gender

After the first dispatch, we see that the men have a negative partial R-indicator, reflecting an underrepresentation compared to the women. After the reminders we see a further increase in the underrepresentativity for the men and, similarly, an increase in the overrepresentation for the women.

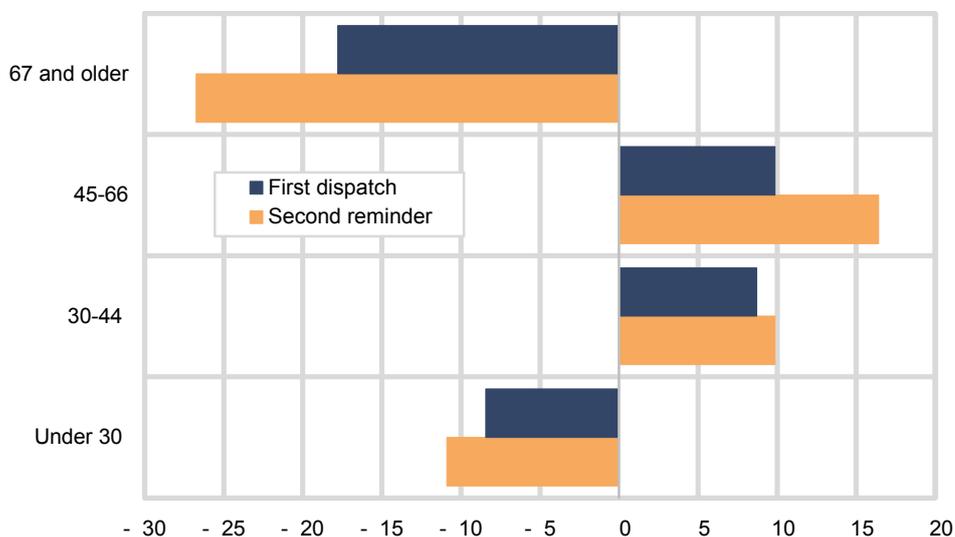
Figure 3.2. The category level indicators for gender, after each dispatch. Multiplied by 1 000



Age

Further along in the data collection process the variable level indicator for age worsens. We see that both the youngest and the oldest categories are underrepresented, while those between 30 and 66 years are overrepresented. Figure 3.3 shows the over and under representativity for the four age categories after the first dispatch and second reminder.

Figure 3.3. The category level indicators for age, after the first and the last dispatch. Multiplied by 1 000



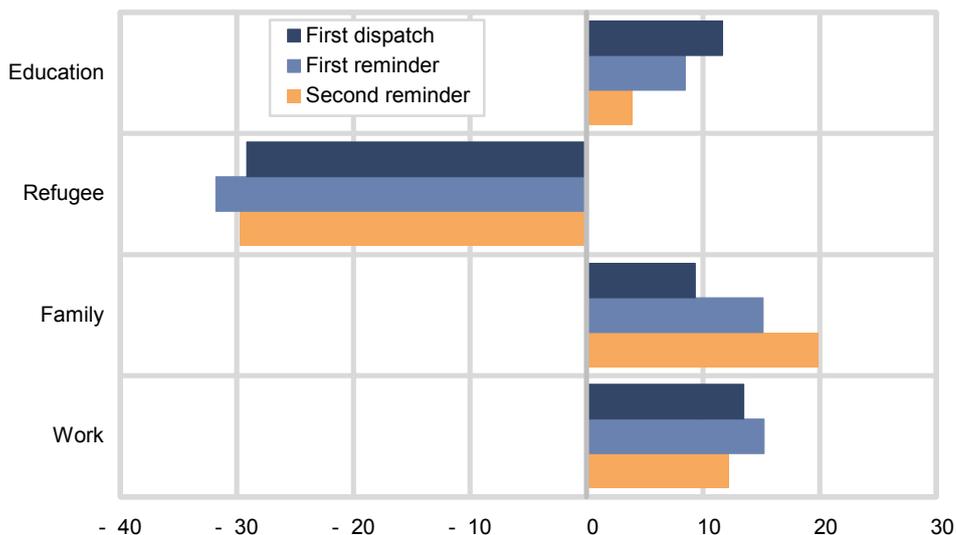
Continent

The variable level indicator for continents is relatively stable during the data collection process. However, when we take a closer look at each continent we find a different pattern. Both Africa and Asia are underrepresented. However, while Africa becomes more and more underrepresented after the two reminders, Asia goes in the other direction and becomes more representative after the reminders. On the other hand, Europe is overrepresented, but stable. The other continents are relatively stable around zero, indicating a good representativity.

Reason for immigration

Figure 3.4 shows that the immigrants with education as reason for immigration become less overrepresented after the reminders, while those who immigrated for family reasons become more and more overrepresented. The work immigrants are also overrepresented, but the representativity remains relatively stable during the data collection process. The refugees also have relatively stable representativity, but are strongly underrepresented. The group with missing reason for immigration is also underrepresented. Because of delays in register information, this group consists of persons who immigrated after 1 January 2011.

Figure 3.4. Categorical level indicators for the main reasons of immigration after each dispatch. Multiplied by 1 000

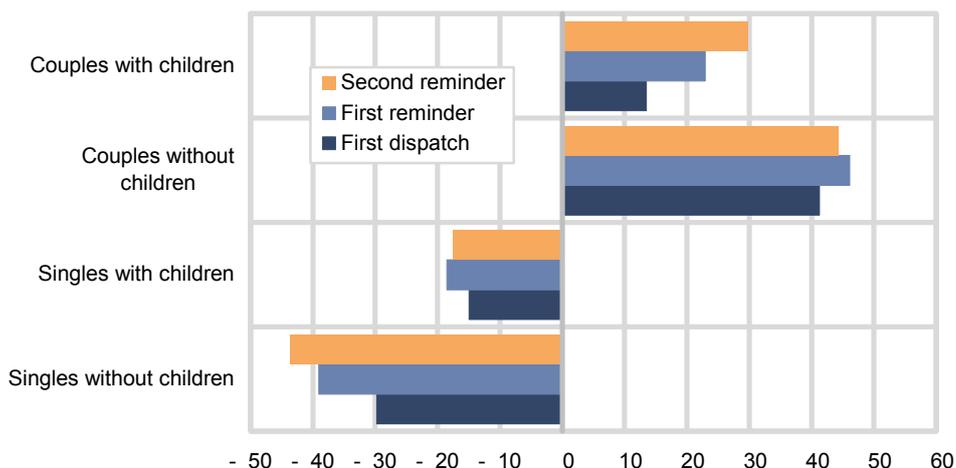


Years living in Norway

The immigrants who have lived in Norway for the shortest and the longest periods, i.e. under two years and five years or more, are underrepresented. However, while the newest immigrants become more underrepresented after the reminders, the opposite is the case for the immigrants who have lived in Norway for five years or more.

Household

The variable level indicator for household is the highest of all the variables and continues to increase after the first reminder and even more after the second reminder.

Figure 3.5. The category level indicator for household types, after each dispatch. Multiplied by 1 000

Couples with and without children are overrepresented, but particularly those without. Those living alone are heavily underrepresented and the underrepresentativity worsens after the reminders. Singles with children are also underrepresented, but not to the same degree. Figure 3.5 shows the category level indicator for the four household types after each dispatch.

Work status

The variable level indicator is also relatively high for work status. Immigrants who are not working³ are underrepresented. The working group is, on the other hand, highly overrepresented. For the missing group, the same applies as for the variable Reason for immigration.

3.3. The level of education

In this paper we have focused on immigrants reporting a high level of education. From the Norwegian education register we know that 28 per cent of persons aged 16 years or older living in Norway are registered with a high level of education (Statistics Norway, 2011). Table 3.3 shows that the percentage of immigrants who reported a high level of education in the questionnaire is considerably higher. The results from the first dispatch give the highest level of education. After the two reminders we see a decline in the percentage of persons with a high level of education, and similarly, an increase in the percentage with less or no education. However, the percentage with a high level of education is still high, even after the second reminder. Since we do not know the education of the non-respondents, we cannot calculate the R-indicator for education level.

Table 3.3. Reported education level among the respondents after each dispatch

	No education	Elementary	Upper secondary	Higher education
After first dispatch	2%	12%	38%	47%
After first reminder	3%	14%	39%	44%
After second reminder	3%	14%	40%	42%

Many of the categories get a very high percentage with a high level of education. For example, after the first dispatch, 53 per cent of the women report having higher education. After the second reminder, the percentage is still very high at 48, and is 11 percentage points higher than the men.

³ According to the employment register from the fourth quarter of 2010.

Figure 3.6. The percentage with a high level of education by gender for each dispatch, reported in the questionnaire

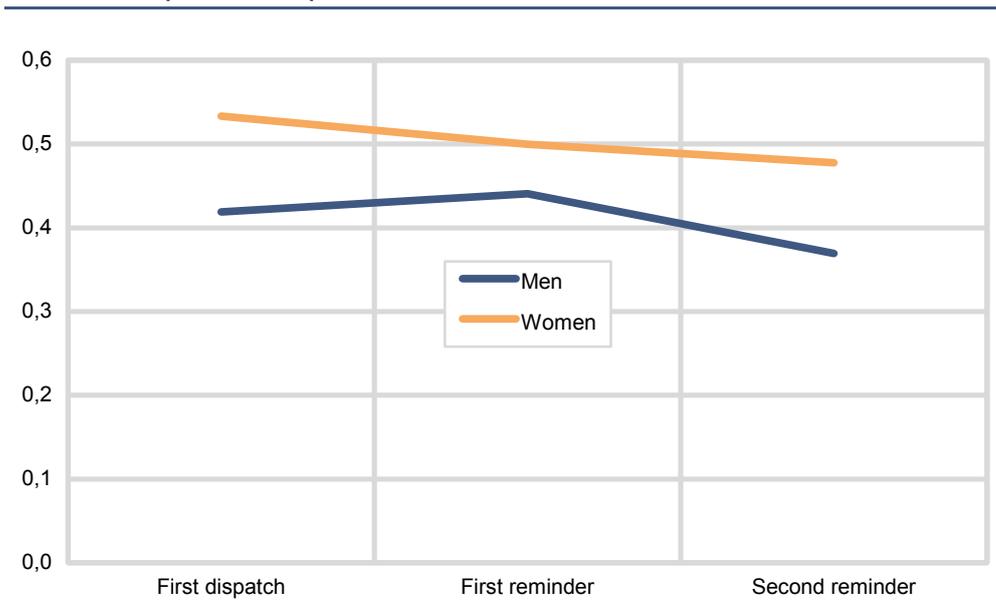
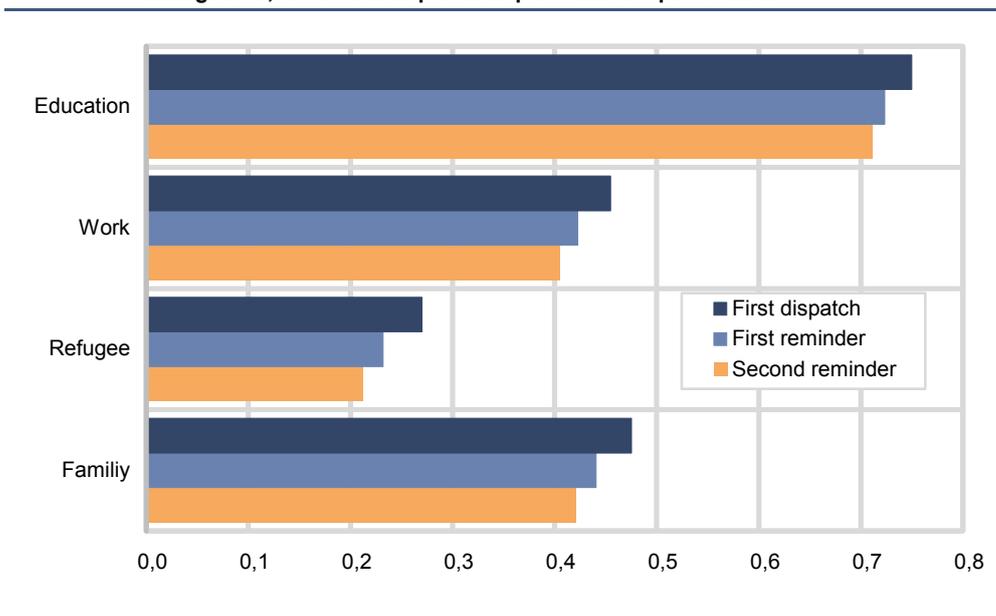


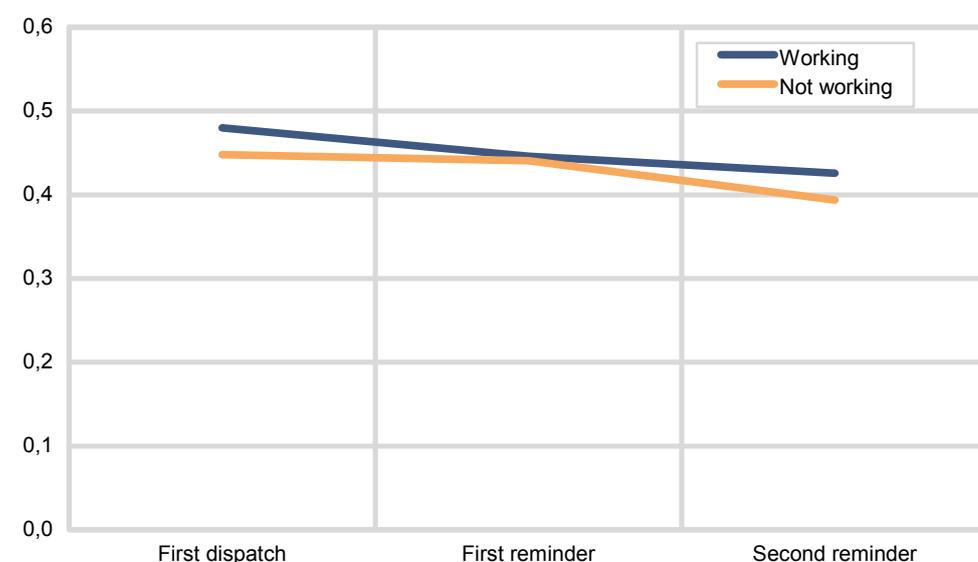
Figure 3.7 shows the percentage with a reported high level of education by reason for immigration after each dispatch with similar patterns; a slightly decreasing percentage with a high level of education.

Figure 3.7. The percentage with a high level of education among the respondents by reason of immigration, after each dispatch. Reported in the questionnaire



We expected to see a correlation between work status and a high level of education. Figure 3.8 shows that the working immigrants have a higher percentage with a high level education, but only slightly. After the second reminder, the immigrants who are not working end up with 39 per cent with a high level of education, while 6 per cent have reported no education.

Figure 3.8. The percentage with a high level of education by work status for each dispatch, reported in the questionnaire



Only the refugee category reports a lower education level than the overall education level in Norway. Immigrants in the oldest age group (67 years and older) report about the same percentage with a high level of education as the overall level in Norway; 29 per cent compared to 28 per cent in the register. The rest of the 29 categories of the auxiliary variables have a considerably higher percentage with a high level of education.

4. Discussion

4.1. Did the reminders improve representativity?

Testing for our seven variables, we see that after the reminders the net sample moves further away from representativity, depending on the Xs in the model. The overrepresented groups after the first dispatch become even more overrepresented after the reminders, and similarly, the underrepresented groups become even more underrepresented after the reminders.

4.2. Did the reminders affect what responses were received in terms of education level?

We assume that many factors can pull the percentage of respondents with a high level of education up but it is difficult to find any factors that pull in the other direction. We suspect that the very small percentage with no education is underreported due to, among other things, illiteracy and language difficulties. The call centre also reported receiving calls from immigrants who did not think we were interested in their lack of or low education. It is reasonable to expect that immigrants with a high level of education are more willing to answer the questionnaire. We know from other surveys in Statistics Norway that people with higher education more frequently answers our questionnaires, e.g. Pedersen (2011) and Pedersen (2012). Furthermore, since this questionnaire asks for the education level, we assume that this effect might be even stronger for this survey. For people with low or no education, specifying their education level may well be a sensitive topic (Dalheim 2002).

The percentage of immigrants with a high level of education decreases after the reminders, but is still high at 42 per cent; 14 percentage points higher than the overall percentage with a high level of education in the Norwegian education

register. The response rate in 1999 was 84 per cent, and 27 per cent reported having completed their higher education (Dalheim, 2002). By sending out reminders, we believe that we were able to reduce the effect of some of the factors described in the section above. The decreasing percentage with a high level of education after the reminders supports this assumption.

5. Conclusion

When considering only the R-indicator, the conclusion seems to be that the representativity for the variables we have tested for worsens after the reminders. The overrepresented groups become even more overrepresented, and similarly, the underrepresented groups become more underrepresented. From this perspective, sending out the reminders seems like a bad idea.

However, we must bear in mind that the main goal was to get as many answers as possible in order to update the Norwegian education register. There is no doubt that the reminders generated more education information than the first dispatch, also from the “hard to get” people through the reminders. However, the response rate for the “not so hard to get” people also seems to increase further, since the representativity worsens for most of the tested variables after each reminder. We believe this percentage is closer to the reality because more people with middle, low or no education respond further into the data collection process. However, after the second reminder, the percentage with a high level of education is still very high.

When we consider the core of this survey, i.e. level of education, it seems that the reminders are indeed useful and necessary regardless of the fact that those with a high level of education seem to be more motivated to respond to the education survey, and that they also respond early in the data collection process. It seems that the representativity for the “hard to get” people with a lower level of education increases after the reminders.

We conclude that sending out reminders worsened the representativity with respect to our seven auxiliary variables. However, with regard to the main goal to fill the gap in the education register, the reminders were necessary. Nevertheless, we believe that the reported percentage with a high level of education is too high to be used for statistical purposes without further adjustments.

References

- Bethlehem, Cobben, Schouten (2011): Handbook of non-response in household surveys. Wiley.
- Cobben and Schouten (2008): An empirical validation of the R-indicator. Discussion paper 08006. Statistics Netherlands.
- Dalheim (2002): En skjemabasert komplettering av registeret over befolkningens høyeste utdanning. Opplysninger om opplæring, skolegang og utdanning 1999. Notater 53/2002.
- Fosen, Kleven, Lagerstrøm, Luiten and Wetzels (2010): Indicators and data collection control. Report from the pilots in Statistics Norway and Statistics Netherlands.
- Pedersen (2011): Undersøkelse blant foreldre med barn som mottok hjelpestønad. Statistics Norway. Notater 37/2011.
- Pedersen (2012): Undersøkelse om ditt nabolag. Statistics Norway 2012. To be published in 2012 in the series Notater by Statistics Norway.
- Schouten and Shlomo (2010): RISQ manual. Tools in SAS and R for the computation of R-indicators and partial R-indicators. Work package 8.
- Schouten and Cobben (2007): R-indexes for the comparison of different fieldwork strategies and data collection modes. Discussion paper 07002, Statistics Netherlands.
- Schouten, Cobben, Bethlehem (2009): Indicators for the representativeness of survey response.
- Schouten, Shlomo and Skinner (2010): Indicators for Representative Response. Paper presented at Q2010, Helsinki, Finland.
- Statistics Norway (2011): Utdanningsstatistikk. Befolkningens utdanningsnivå, 1 October 2010. Stadig økende andel med høyere utdanning. <http://www.ssb.no/utniv/>. Downloaded 25 April 2012.
- Witt (2010): Estimating the R-indicator, Its Standard Error and Other Related Statistics with SAS and SUDAAN.

List of figures

3.1. The R-indicators for the seven auxiliary variables	8
3.2. The category level indicators for gender, after each dispatch. Multiplied by 1 000.....	10
3.3. The category level indicators for age, after the first and the last dispatch. Multiplied by 1 000	10
3.4. Categorical level indicators for the main reasons of immigration after each dispatch. Multiplied by 1 000	11
3.5. The category level indicator for household types, after each dispatch. Multiplied by 1 000	12
3.6. The percentage with a high level of education by gender for each dispatch, reported in the questionnaire.....	13
3.7. The percentage with a high level of education among the respondents by reason of immigration, after each dispatch. Reported in the questionnaire.....	13
3.8. The percentage with a high level of education by work status for each dispatch, reported in the questionnaire.....	14

List of tables

1.1. Key figures from the survey.....	6
3.1. The overall R-indicator after each dispatch	8
3.2. Partial R-indicators: the variable level indicators (in bold) and the category level indicators for each of the seven auxiliary variables. Multiplied by 1 000	9
3.3. Reported education level among the respondents after each dispatch	12

B Returadresse:
Statistisk sentralbyrå
NO-2225 Kongsvinger

Statistics Norway

Oslo:

PO Box 8131 Dept

NO-0033 Oslo

Telephone: + 47 21 09 00 00

Telefax: + 47 21 09 00 40

Kongsvinger:

NO-2225 Kongsvinger

Telephone: + 47 62 88 50 00

Telefax: + 47 62 88 50 30

E-mail: ssb@ssb.no

Internet: www.ssb.no

ISBN 978-82-537-8403-8 Printed version

ISBN 978-82-537-8404-5 Electronic version

ISSN 1891-5906

ISBN 978-82-537-8403-8



9 788253 784038



Statistisk sentralbyrå
Statistics Norway