1351.0.55.039

Australian Bureau of Statistics

Research Paper

Investigating the Discrepancy between Measured and Self-Reported BMI in the National Health Survey

www.abs.gov.au

New Issue

Research Paper

Investigating the Discrepancy between Measured and Self-Reported BMI in the National Health Survey

Tim Ayre, Jason Wong and Anil Kumar

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 16 MAR 2012

ABS Catalogue no. 1351.0.55.039

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Ruel Abello, Analytical Services Branch, on Canberra (02) 6252 6307 or email <analytical.services@abs.gov.au>.

CONTENTS

ABSTRACT 1

1. INTRODUCTION 2

2. PAST STUDIES ON BMI MISREPORTING 3

3. DATA 5

4. TRENDS IN BMI, HEIGHT AND WEIGHT AND SELECTED FACTORS ASSOCIATED WITH MISREPORTING 7
   - 4.1 Reported *vs* measured BMI 7
   - 4.2 Reported *vs* measured height 8
   - 4.3 Reported *vs* measured weight 9
   - 4.4 Relationship between age and misreporting 10
   - 4.5 Relationship between reporting error against reported values 11
   - 4.6 Relationship between self-reported health status and misreporting 14

5. MODELLING THE DISCREPANCIES BETWEEN REPORTED AND MEASURED DATA 15
   - 5.1 Linear regressions 16
   - 5.2 Semi-parametric regressions 23

6. ADJUSTING BMI CATEGORY ESTIMATES 27
   - 6.1 Assessing the accuracy of corrected estimates 27
   - 6.2 BMI category estimates for the population 30

7. CONCLUSION AND RECOMMENDATIONS 31

REFERENCES 33

APPENDIXES

A. EXTENDED MODELS WITH AGE GROUPS, USING POOLED 1995 AND 2007–08 DATA 34

B. ESTIMATED SMOOTH TERMS FROM GAM MODELS 37

C. GAM MODELS WITH AGE GROUPS 39

D. ADJUSTED BMI CATEGORY ESTIMATES 42

E. BMI CATEGORY CLASSIFICATION 43

INVESTIGATING THE DISCREPANCY BETWEEN MEASURED AND SELF-REPORTED BMI IN THE NATIONAL HEALTH SURVEY

Tim Ayre, Jason Wong and Anil Kumar
Analytical Services Branch

ABSTRACT

Body Mass Index (BMI) is commonly used to measure the prevalence of obesity in populations for health research. Population surveys often ask respondents to report their height and weight, rather than taking physical measurements. Previous research has shown that the discrepancies between self-reported and measured values of height and weight can lead to inaccurate estimates of the population BMI distribution. The accuracy of estimates derived from self-reported BMI data can potentially be improved by adjusting the self-reported values to account for these reporting biases.

In this paper we investigate the reporting errors in height, weight and BMI of Australian adults using the National Nutrition Survey (NNS) 1995 and National Health Survey (NHS) 2007–08. Both surveys collected measured and self-reported height and weight. Linear and semi-parametric regressions are used to adjust self-reported BMI in NHS 2007–08, and the resulting BMI distributions are compared with the distributions of measured and self-reported BMI.

The results confirm that, on average, respondents overestimate their height and underestimate their weight and BMI. However, the magnitude of the misreporting was significantly smaller in the NHS 2007–08 than in NNS 1995. Adjusting self-reported BMI is found to provide significantly more accurate estimates of the distribution of BMI than using self-reported BMI directly. However, more data are necessary to assess whether the levels and pattern of misreporting have stabilised such that actual BMI can be accurately predicted from self-reported data using modelling.

1. INTRODUCTION

Rising obesity in the Australian population is a major health concern. Obesity is a key indicator of risk for conditions such as heart disease, high blood pressure and diabetes, particularly when linked with other lifestyle factors such as lack of exercise and smoking. It is important that reliable and accurate data on obesity are available to authorities to monitor the magnitude of the problem and assess the effectiveness of interventions.

The Body Mass Index (BMI), based on a person's height and weight, is commonly used to measure obesity. To estimate the BMI of respondents, population surveys often collect self-reported height and weight data rather than measured data, which are more complex and costly to collect. However, Australian and overseas studies suggest there can be discrepancies between the two different measures. It has been found that, for example, there is a tendency for people to over-report their height and under-report their weight, leading to an underestimation of BMI. As a result it is possible that using obesity estimates based on self-reported measures could lead to significant underestimation of the true size of the problem.

The objective of this paper is to investigate misreporting of BMI among Australian adults using data from two ABS National Health Surveys (NHS). The National Nutrition Survey (NNS) 1995[1] and the 2007–08 NHS have both reported and measured data. By looking at these we identify several factors associated with misreporting. Several alternative models were developed to examine the difference between reported and measured data, and to assess whether self-reported figures can be adjusted to give more accurate estimates of the true BMI distribution. Analysis of misreporting is undertaken using individual-year and pooled data to examine whether misreporting has increased or decreased between the two survey periods. We also examine factors associated with height and weight misreporting separately.

The remainder of this paper is organised as follows. Section 2 presents a brief review of the literature on misreporting of BMI and methods for correcting this discrepancy. Section 3 describes the data used in this study. Section 4 presents some descriptive statistics on the trends in BMI, height and weight and selected factors associated with misreporting using data from the 1995 and 2007–08 surveys. Section 5 presents and discusses model results from linear and semi-parametric modelling. Section 6 presents adjusted BMI distributions estimated using the fitted models and compares them to the distributions estimated from measured and reported data. Section 7 concludes the paper and makes recommendations.

[1] The NNS 1995 is a subsample of NHS 1995 that contains extra variables.

2. PAST STUDIES ON BMI MISREPORTING

BMI is an important indicator of a person's health status and is closely linked to chronic diseases such as diabetes and heart disease. There is a wide range of literature on the study of BMI, including misreporting of BMI.

McAdams *et al.* (2007), using a study of US adults, found that while self-reported BMI was sufficiently accurate for calculating correlations with disease biomarkers, there was evidence of under-reporting of BMI, especially among overweight and obese persons. Females under-reported their weight by 1.47 kg on average, but males over-reported their weight by an average of 0.37 kg. Height was over-reported by both males and females (by 1.16 cm and 0.37 cm respectively). Average under-reporting of BMI was 0.22 kg/m^2 for males and 0.67 kg/m^2 for females, but for obese persons (measured BMI greater than or equal to 30) average under-reporting increased to 1.36 kg/m^2 for males and 2.09 kg/m^2 for females.

Wang *et al.* (2002) investigated the differences between self-reported and measured height, weight and BMI distributions for Australian adolescents aged 15–19 from the 1995 National Nutrition Survey (NNS). Overweight and obese adolescents were found to have much greater downward bias in their self-reported weight than their normal or underweight counterparts. They did not, however, find any statistically significant difference in misreporting of weight and BMI between male and female adolescents, in contrast to findings from other studies based on adult respondents.

Hayes *et al.* (2008) used the 1995 NNS to estimate correction equations to adjust self-reported height, weight and BMI for persons aged twenty years and over. Two sets of correction equations, both fitted separately for males and females, were estimated. The first set, the simple models, included self-reported values of the variable of interest as the only explanatory variable. The second set, denoted as the extended models, included income decile, age category, smoking status, age left school, marital status and employment status as additional explanatory variables in addition to the self-reported value. The results from these regressions can then be used to adjust the self-reported values of individuals for whom measured values were not recorded. The paper reported that the adjusted BMI from both the simple and extended models provided improved BMI category estimates for the population. However, the extended models provided only small improvements to the estimates when compared to the simple models. Another finding was that deriving the BMI from adjusted height and weight was slightly more accurate than simply adjusting the BMI derived from self-reported figures directly using the correction equation for BMI.

Dauphinot *et al.* (2009) used receiver operating characteristic curves to determine an alternative obesity cut-off when using self-reported BMI data. This approach adjusts the BMI threshold used to define obesity, taking account of false positive and false negative rates, rather than adjusting each individual's BMI estimate. It therefore does

not take account of individual characteristics other than reported BMI, although different cut-offs could be determined for different groups (e.g. males and females). Using a sample of Swiss adults aged between 34 and 75 taken between 1993 and 2004, Dauphinot *et al.* (2009) calculated an alternative obesity cut-off of 29.2 kg/m^2 for both males and females. This cut-off was validated using a 2002–03 survey of French adults aged 18 or over. The accuracy of adjusted obesity estimates from self-reported figures greatly improved, implying that adjustment is important when providing statistics on BMI from self-reported figures.

In conclusion, the literature cited in this section is consistent in the sense that under-reporting of BMI is prevalent, and self-reported BMI must be adjusted in order to obtain more reliable statistics. In this paper, we examine BMI misreporting in the Australian context using data from two surveys, including the most recent National Health Survey. We explore alternative correction models that could be used to adjust reported data to provide more accurate estimates of obesity prevalence.

3. DATA

This paper uses data from two surveys: the 1995 National Nutrition Survey (NNS), which was conducted on a subsample of the 1995 NHS; and the 2007–08 National Health Survey (NHS) (ABS 1995, 2007–08). The NNS 1995 was conducted from February 1995 to March 1996, with height and weight measurements obtained from a subsample of persons in the NHS 1995 (ABS 1995). The NHS 2007–08, which was conducted from August 2007 to July 2008, collected measured height and weight data from approximately 70 per cent of respondents in the survey. Both surveys also collected information on a range of demographic and health-related characteristics and activities.

The total sample sizes were 13,858 persons for NNS 1995 and 20,788 persons for NHS 2007–08. For the purposes of this analysis we excluded the following: persons under 18 years old[2]; those observations with missing measured or reported height or weight data; and some extreme outliers.[3] These exclusions lead to a sample of 9,805 persons for NNS 1995 and 9,271 persons for NHS 2007–08.

Respondents in the two surveys, except those who were pregnant, were asked to give an approximate estimate of their height (without shoes) and weight. Respondents could answer in metric or imperial units. When respondents gave an imprecise response (e.g. when 'around' was used in the response), they were prompted to give a more precise answer. Responses given in imperial units were converted to metric units, and self-reported height and weight were rounded to the nearest 1 cm and 1 kg respectively.

Measuring of weight and height was voluntary for respondents. Height was measured using a stadiometer and weight using a digital scale. Interviewers were trained for taking the physical measurements. A distinction between the two surveys was that measured height and weight were rounded to the nearest 0.1 cm and 0.1 kg respectively for NNS 1995, but to the nearest 1 cm and 1 kg respectively for NHS 2007–08. Furthermore, for NNS 1995 height was measured without shoes and weight was measured with only a single layer of light clothing and without shoes, while for the NHS 2007–08 these requirements were voluntary and some respondents may not have complied with them. These differences, however, are not expected to significantly affect data comparability across the two surveys.

---

2 There are two reasons for confining our analysis to adults or 18+. First, there is literature that shows that reporting behaviour of adolescents is different from that of adults. Second, there was a large proportion of missing observations in the children's sample. For instance, 92% of the children's sample had either reported or measured data missing in 2007-08.

3 These are defined here as observations where the difference between measured and reported height/weight is more than 4 standard deviations higher or lower than the mean.

The BMI measure in the dataset is computed from weight and height data using the following standard formula:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (m)}^2}$$

In this paper, self-reported BMI refers to the value derived from self-reported height and self-reported weight, not a directly reported value. Measured BMI refers to the value derived from measured height and weight.

For the regression modelling, the datasets from both surveys were combined to create a pooled dataset. Pooling the two datasets increases the sample size and hence reduces the sampling errors. However, this approach implicitly assumes that the coefficients of the explanatory variables are the same for the two time periods.

The inclusion of period effect in the models also allows direct evaluation of the change in expected misreporting between the two surveys holding other variables constant. Any period effects that show up will fall into two categories: real-world effects due to the time periods in which the surveys were conducted; and those due to changes in the surveys themselves, such as differences in the sample design, scope/coverage, collection methods, questionnaire wording, etc.. The two surveys appear to be comparable in terms of overall design and collection of the variables that are used in the regressions. Given the similarities between the two surveys, we can assume that any large differences between the two surveys mainly reflect differences between the two time periods.

4. TRENDS IN BMI, HEIGHT AND WEIGHT AND SELECTED FACTORS ASSOCIATED WITH MISREPORTING

In this section we present some descriptive statistics on the trends in BMI, height and weight and selected factors associated with misreporting using data from the 1995 and 2007–08 surveys.

4.1 Reported *vs* measured BMI

Table 4.1 below shows the distribution of reported and measured BMI by sex for the two survey periods, for persons aged 18 and above. The data includes only those persons for whom both reported and measured data were available and it excludes outliers. The results show that the proportion of persons reporting BMI in the 'normal' category fell and the proportion of persons reporting overweight/obese rose from 1995 to 2007–08.[4] This trend is observed for both males and females over this period. For males only the obese category increased, with more or less no change in the overweight category. For females both overweight and obese categories recorded increases (with a larger rise in the obese category).

4.1 Proportion of persons by reported or measured BMI, by sex, 1995 and 2007–08

	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
	Reported BMI			
Underweight	1.1	4.8	1.1	3.7
Normal	45.4	58.1	35.8	47.9
Overweight	41.9	25.7	41.8	28.8
Obese	11.7	11.4	21.3	19.7
	Measured BMI			
Underweight	0.5	2.3	0.9	2.5
Normal	33.8	49.1	30.2	42.7
Overweight	46.8	31.1	43.4	31.4
Obese	18.8	17.4	25.6	23.4

Similar trends are observed in the measured data. While the discrepancy in the proportion measured as overweight/obese fell between the two periods, a significant difference between reported and measured data still remained in 2007–08 (63.1% *vs* 69.0% for males and 48.5% *vs* 54.8% for females). This difference is due only to misreporting of BMI, and not selection bias, as the same records are used to calculate both the measured and reported estimates.

[4] Underweight is defined as BMI less than 18.5, 18.5-25 for normal, 25-30 for overweight and greater than 30 for obese.

4.2 Reported *vs* measured height

Height misreporting (difference between reported and measured height) appears to have declined over time. As shown in table 4.2, in 1995 only 12.0% of males and 15.7% of females reported height correctly to the nearest centimetre (i.e. height discrepancy of zero), but this proportion rose to around a quarter in 2007–08 for both males and females. The proportion of persons with a reported height discrepancy of between –1 and 1 cm rose from 34.6% to 50.3% for males and from 41.5% to 51.6% for females. The proportion over-reporting height (i.e. discrepancy greater than 0) for both males and females has fallen over time, from 71.8% to 51.2% for males and 58.6% to 48.4% for females. The proportion of males under-reporting height (i.e. discrepancy less than 0) has increased from 16.2% to 23.2%; the increase for females was much smaller (from 25.8% to 26.8%). The mean discrepancy in height has fallen from just over 2 cm to 1 cm for males and from 1.3 cm to 0.8 cm for females over this period.

4.2 Distribution of persons by height discrepancy, by sex, 1995 and 2007–08

	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
Height discrepancy (cm) (reported minus measured)	%	%	%	%
–15 to –10 cm	0.2	0.2	0.4	0.8
–9 to –8	0.4	0.4	0.8	1.0
–7 to –5	1.0	1.7	2.1	2.3
–4	1.0	1.8	1.5	1.6
–3	1.8	3.7	2.9	3.0
–2	3.9	7.7	5.3	6.2
–1	7.8	10.2	10.3	12.0
0	12.0	15.7	25.6	24.8
1	14.8	15.7	14.4	14.9
2	16.0	13.9	11.5	12.2
3	14.7	9.9	9.3	7.6
4	9.6	6.4	6.0	4.5
5 to 7	13.2	9.0	7.7	6.4
8 to 9	2.2	2.0	1.5	1.5
10 to 15 cm	1.4	1.7	0.9	1.4
–1 to 1 cm	34.6	41.5	50.3	51.6
–2 to 2 cm	54.5	63.2	67.0	70.0
<0 cm	16.2	25.8	23.2	26.8
>0 cm	71.8	58.6	51.2	48.4
Mean difference in height	2.00 cm	1.27 cm	0.97 cm	0.75 cm

4.3 Reported *vs* measured weight

Weight misreporting also appears to have declined over time. In 1995 only 10.9% of males and 10.5% of females reported weight correctly to the nearest kilogram (i.e. weight discrepancy of zero) but these proportions rose to 23% and 24% respectively in 2007–08. The proportion of persons with a reported weight discrepancy of between –1 and 1 kg has risen from around 30% for both males and females in 1995, to 45.9% and 50.4% respectively in 2007–08. Although there has been a larger decline in the proportion of females under-reporting their weight over time relative to males, a larger proportion of females still under-reported their weight in 2007–08 relative to males (60% *vs* 54%). The mean discrepancy in weight fell from –1.8 kg to –1.1 kg for males and from –2.6 kg to –1.4 kg for females over this period.

4.3 Distribution of persons by weight discrepancy, by sex, 1995 and 2007–08

	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
Weight discrepancy (kg) (reported minus measured)	%	%	%	%
<–15 kg	0.3	0.7	0.0	0.0
–15 to –11	2.3	1.9	1.3	1.3
–10 to –6	11.4	12.2	6.5	5.3
–5 to –4	14.7	17.7	10.0	10.0
–3	10.3	13.5	8.9	10.9
–2	12.9	16.0	12.5	14.3
–1	12.9	14.9	14.7	18.1
0	10.9	10.5	23.0	24.1
1	8.3	5.7	8.1	8.2
2	5.9	3.3	5.8	3.6
3	3.9	1.4	3.2	1.7
4 to 5	3.6	1.5	3.2	1.5
6 to 10	2.2	0.6	2.4	1.0
11 to 15	0.4	0.1	0.4	0.1
> 15 kg	0.0	0.0	0.0	0.0
–1 to 1 kg	32.1	31.0	45.9	50.4
–2 to 2 kg	50.9	50.3	64.1	68.3
<0 kg	64.7	76.9	53.9	59.8
>0 kg	24.4	12.6	23.2	16.1
Mean difference in weight	–1.82 kg	–2.58 kg	–1.08 kg	–1.41 kg

4.4 Relationship between age and misreporting

Misreporting appears to be related to age, with height-reporting discrepancies appearing to increase among older persons, especially those aged 55 years and above. This relationship is observed for both males and females, and in both periods, although it was less pronounced in 2007–08.

Tables 4.4–4.6 show the percentage of records that report height, weight and BMI within specified benchmarks by age group. The proportion of observations having reporting errors less than the benchmarks for height (± 3cm) and BMI (± 1) decreases as age increases. This suggests that older individuals on average have higher reporting errors for height and BMI. However, no such trend is observed for weight (± 3 kg).

4.4 Proportion reporting height within 3cm of actual height, by age group

	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
Age group				
18–24 years	66.4	75.1	80.5	84.1
25–44 years	72.8	76.7	84.6	83.8
45–54 years	71.5	81.4	81.6	85.2
55–64 years	58.1	70.9	79.7	82.9
65–74 years	42.3	47.1	70.5	73.7
75+ years	28.5	32.0	55.9	59.3

4.5 Proportion reporting weight within 3kg of actual weight, by age group

	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
Age group				
18–24 years	62.3	62.5	77.7	82.8
25–44 years	61.8	61.0	78.8	84.1
45–54 years	54.1	56.9	75.5	79.3
55–64 years	57.8	58.7	75.4	80.4
65–74 years	59.4	59.5	74.1	78.5
75+ years	60.1	61.7	68.7	71.8

4.6 Proportion with BMI within 1 kg/m^2 of actual BMI, by age group

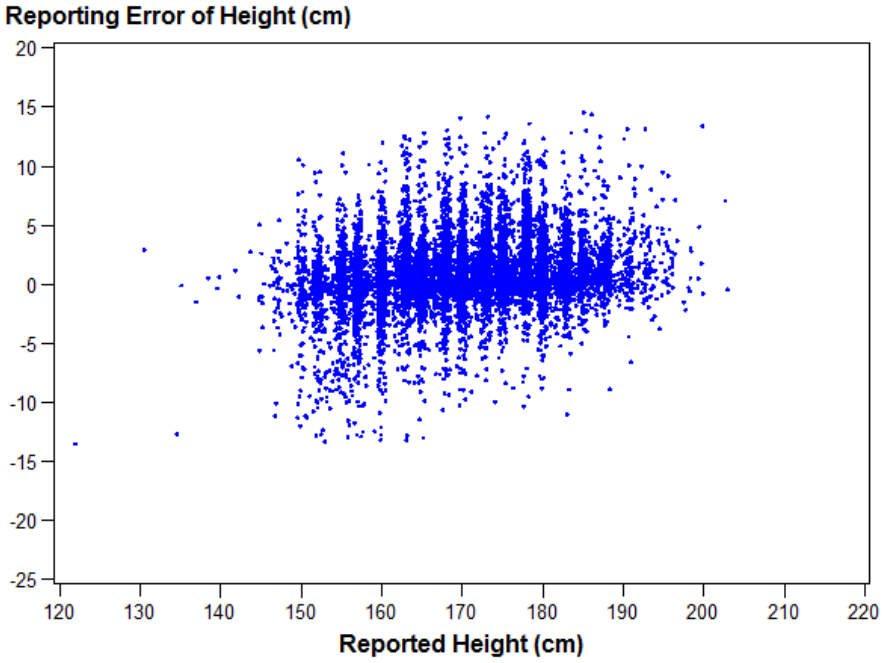
	1995		*2007–08*	
	Male	*Female*	*Male*	*Female*
Age group				
18–24 years	50.7	43.6	65.1	56.2
25–44 years	48.6	44.6	62.6	60.6
45–54 years	38.0	43.4	60.0	55.4
55–64 years	34.4	32.7	49.7	52.4
65–74 years	27.2	23.3	45.5	44.1
75+ years	25.4	16.5	36.2	36.6

4.5 Relationship between reporting error against reported values

Figures 4.7–4.12 show the relationship between reporting errors and the reported value of the measure of interest.

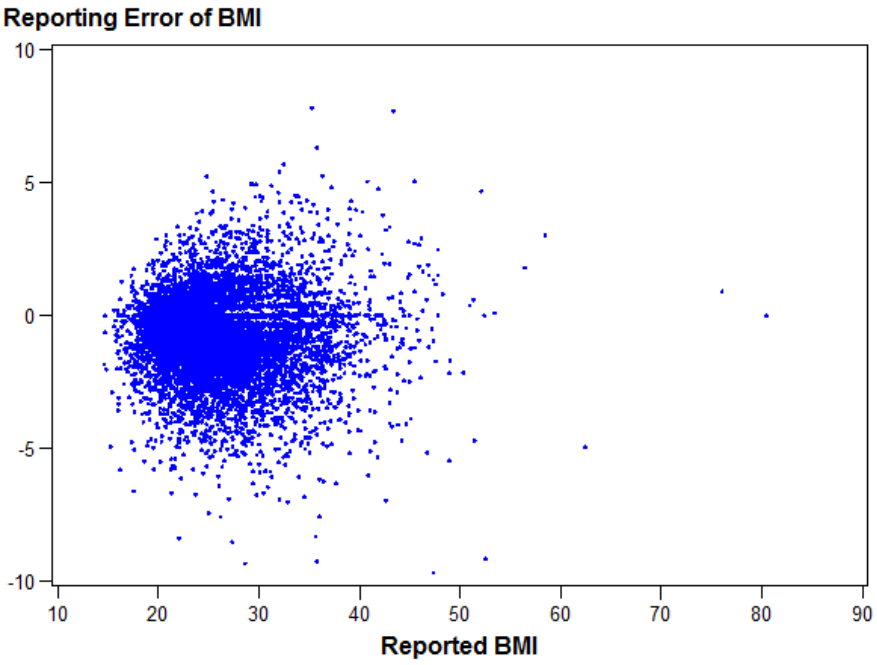
For weight and BMI, figures 4.11–4.12 suggest that as reported weight and reported BMI increase, the expected proportion of observations reporting their weight and BMI within the benchmark decrease respectively. Regarding height (figure 4.10), the same trend is found for females, but for males the accuracy of reported height does not appear to be related to the value of reported height, other than those with reported height between 155 cm and 160 cm having a relatively lower proportion of observations reporting their height within 3 cm of their measured height.

4.7 Reporting error of height *vs* reported height, 2007–08



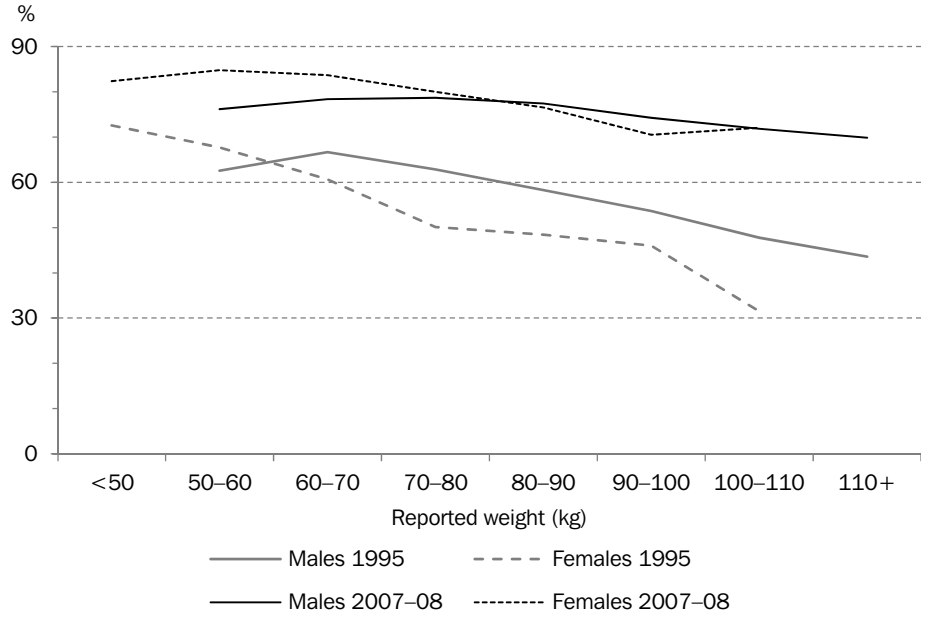
4.8 Reporting error of weight *vs* reported weight, 2007–08

4.9 Reporting error of BMI *vs* reported BMI, 2007–08

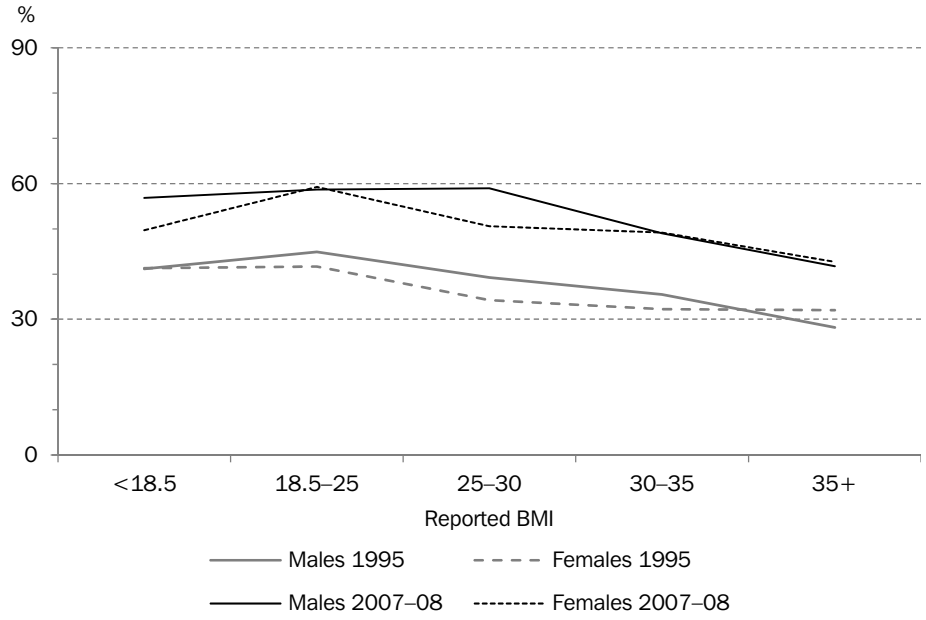


4.10 Percentage of observations reporting height within 3cm, by reported height

4.11 Percentage of observations reporting weight within 3kg, by reported weight



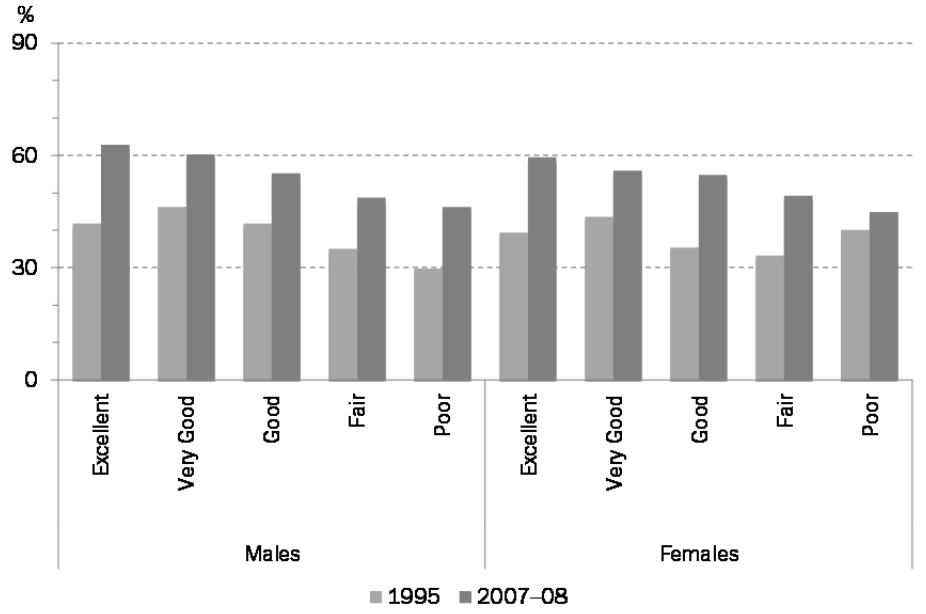
4.12 Percentage of observations reporting BMI within 1 kg/m^2, by reported BMI



4.6 Relationship between self-reported health status and misreporting

Figure 4.13 below shows the relationship between self-reported health status and the proportion of observations reporting BMI within 1 kg/m^2. Persons with poorer health are more likely to misreport more than those with better health. This relationship appears stronger for men than women, and is present in both time periods.

4.13 Proportion with BMI within 1 kg/m^2 of actual BMI, by self-reported health status



5. MODELLING THE DISCREPANCIES BETWEEN REPORTED AND MEASURED DATA

Section 4 showed that self-reported BMI figures are biased downwards, and that the size of the bias depends on various factors. In this section we develop several regression models to examine the feasibility of adjusting self-reported BMI figures for reporting bias.

The regressions model the discrepancy between reported and measured data as a function of relevant explanatory variables. Both linear and semi-parametric regressions are conducted to examine this relationship. The response variable for each regression is the reporting error (i.e. reported minus measured) for the relevant measure (i.e. height, weight or BMI). Such regressions are known as correction equations.

The regressions were estimated using the pooled dataset of 1995 and 2007–08.[5] A survey period indicator was included in all the models to take account of the differences in average misreporting levels between the two periods.

Separate models were estimated for males and females, as the analysis in the previous section showed that the relationships between the reporting discrepancies and several variables varied by sex.

In all of the models, the variables of height, weight, BMI and age were re-centred by subtracting the approximate mean values of these variables[6], as follows:

- Height was re-centred by subtracting 170 cm.
- Weight was re-centred by subtracting 75 kg.
- BMI was re-centred by subtracting 25 kg/m^2.
- Age was re-centred by subtracting 45 years.

Mean-centring of these continuous variables is done here to make the base case represent a roughly typical person, so the intercepts can be usefully interpreted[7] and

---

5 Regressions were also run on the two datasets separately, but the results were largely consistent with the results from the pooled dataset.

6 These values are not the exact means for these variables, but rounded values close to their means. The mean values for reported height, reported weight, reported BMI and age in the pooled dataset were 170 cm, 74 kg, 26 kg/m^2 and 47 years respectively.

7 The intercept in a linear model represents the expected value of the dependent variable with a value of zero for all the explanatory variables. Such observations are often referred to as the base case. In some cases the base case might not make sense (e.g. height being equal to zero), and so the intercept does not necessarily have any meaningful interpretation. By centring the explanatory variables to their mean values, the intercept can be interpreted as the expected value of dependent variable for all the explanatory variables set to their mean values, rather than zero.

compared between different models. This re-centring was done only for ease of interpretation and does not affect the fitted models.[8]

5.1 Linear regressions

Ordinary Least Squares (OLS) regressions were conducted separately for misreporting of height, weight and BMI by sex to examine the relationship between misreporting of each attribute and an individual's characteristics:

$$y_{\text{rep},i} - y_{\text{act},i} = \beta_0 + \sum_k \beta_k X_{ki} + \varepsilon_i$$

where:

$y_{\text{rep},i} - y_{\text{act},i}$ is the reporting discrepancy for individual i;

X_{ki} is the value of the k-th characteristic for individual i;

β_0 is the intercept;

β_k is the coefficient of the k-th characteristic; and

ε_i is the random error for individual i, assumed to be independently and identically distributed for all individuals.

For the linear regressions we considered two sets of models, denoted Simple OLS and Extended OLS. The simple OLS models contain only a few explanatory variables. These are reported height, weight and BMI, sex, age, age-squared[9], self-assessed body weight (acceptable/underweight *vs* overweight) and the survey period indicator. Only those variables that were found to be statistically significant were retained in the model and insignificant variables were dropped.

The Extended OLS models add statistically significant socio-economic variables and other relevant risk factors to the Simple OLS models. These extra variables were self-assessed body weight being underweight, self-assessed health, smoking status, labour force status, and country of birth (Australia *vs* overseas). Interaction terms of some of these extra variables with the Period variable are also included to allow for different effects between the two time periods for these factors.

---

8 The slope coefficients for linear variables remain the same after re-centring, but the coefficients for variables that have interaction or squared terms will be changed by the re-centring. However the models are still equivalent when re-centred.

9 We also considered age as a categorical variable in the Simple OLS models but did not find any significant improvement in the model fit. These results are not presented here but age as a categorical variable is considered in the Extended OLS models in Section 5.1.2.

The reason for having both simple and extended models is to examine whether simpler models can provide a good enough approximation of the correction error or whether additional variables are required to capture this phenomenon better. Note that while ideally the selection of explanatory variables should be based on theory and what best captures the relationship between the dependent and the explanatory variables, for the purposes of this analysis we had to confine the modelling to what variables were available in the dataset and what other researchers have used.

5.1.1 OLS results

Tables 5.1–5.3 show the coefficients, adjusted R-squared (R^2) statistics and Root Mean Squared Error statistics (RMSE) of the OLS models for height, weight and BMI misreporting respectively. The base categories in the regressions were: male; those with acceptable/underweight self-assessed body weight; and 1995. Variables that were not statistically significant were dropped.

Interpretation of model results

The coefficients in each model indicate the impact of the explanatory variables on the reporting error for the relevant attribute, holding all other explanatory variables constant. A positive coefficient indicates that the greater the value of the explanatory variable, the greater the overestimation or the smaller the underestimation, depending on the direction of the expected misreporting. On the other hand, a negative coefficient indicates that the greater the value in the explanatory variable, the smaller the overestimation or the greater the underestimation.

For example in table 5.1, the coefficient of –0.0293 for Reported Weight for males can be interpreted as the expected overestimation in height decreases by 0.0293 cm for each 1 kg increase in reported weight, holding other variables constant, while in table 5.2 the coefficient of 0.0254 for Reported Weight for males can be interpreted as expected underestimation in weight decreases by 0.0254 kg for each 1 kg increase in reported weight.

The marginal effect of an explanatory variable with a quadratic term depends on the signs and values of the coefficients for the variable. For example, in the simple models in table 5.1, the coefficients for Reported Height and Reported Height^2 for males indicate that for each 1 cm increase in reported height, over-reporting in height increases by (0.1849 – 0.0072 × Reported Height), holding other variables constant. This means that over-reporting increases at a decreasing rate until reported height reaches 196 cm where it then starts to decrease. For Age, where both the linear and quadratic terms are positive, over-reporting in height for males increases by (0.0444 + 0.002 × Age) for each 1 year increase in age, which indicates that over-reporting in height is higher for older individuals and rises at an increasing rate.

The coefficients for the Extended OLS are interpreted in the same way as in the Simple OLS models, except for interaction terms between survey year and the additional variables that were added, where statistically significant for either males or females. The addition of interaction terms does not affect those who are in the reference groups for the variables that are involved in the interaction terms. However, for those who are not in the reference groups, a constant is added to the expected reporting error in 2007–08, depending on the coefficients of the interaction terms and the characteristics of the individuals. For example, for a male who is unemployed, the expected reporting error in height in 2007–08 is 0.053 cm (0.3459–0.2929, refer to table 5.1) less than an individual who is employed. The interaction terms were included to allow the the impact of the specific variables on misreporting to vary over time.

Height misreporting

Table 5.1 shows the fitted models for height misreporting. The positive intercepts imply that individuals in the base case (i.e. those with height, weight, BMI and age close to the population mean values) overestimate their height on average. The tables indicate that females in the base case have a greater overestimation of height than males in the base case. The positive and negative coefficients for males for reported height and (reported height)-squared, respectively, imply that the reporting error in height increases with height but at a decreasing rate, with expected overestimation reaching a maximum at a certain value of reported height[10] before declining. However for females, the coefficient for (reported height)-squared is close to zero and the trend of overestimation against reported height is close to linear against height. The results also indicate that average over-reporting of height was lower in 2007–08 than in 1995.

The adjusted R^2 values indicate the models do not have high goodness-of-fit. Only between 20% and 26% of the variation in reporting error of height is explained by the models. The R^2 statistics suggest that there is a better fit for females than males.

[10] These values were 196 and 197 cm for Simple and Extended OLS respectively.

5.1 OLS model of height misreporting, using pooled 1995 and 2007–08 data

	Male		*Female*	
	Simple	*Extended*	*Simple*	*Extended*
Dependent variable: Reported minus measured height, in cm				
Intercept	0.9884 ***	0.7962 ***	1.9189 ***	1.0516 ***
Period[+]	–1.0820 ***	–0.9141 ***	–0.7054 ***	–0.2509 ***
Reported height	0.1849 ***	0.1940 ***	0.1747 ***	0.1875 ***
(Reported height)^2	–0.0036 ***	–0.0036 ***	0.0000	–0.0001
Reported weight	–0.0293 ***	–0.0399 ***	–0.0133 ***	–0.0278 ***
Age	0.0444 ***	0.0402 ***	0.0447 ***	0.0357 ***
Age^2	0.0010 ***	0.0010 ***	0.0017 ***	0.0014 ***
Self-assessed weight				
Underweight		–0.3485 ***		0.0802
Acceptable (=reference)				
Overweight		0.3645 ***		0.6003 ***
Self-assessed health				
Excellent, very good or good (=reference)				
Fair / poor health		0.2508 ***		0.4356 ***
Smoking status				
Current smoker (=reference)				
Never smoked		–0.1799 ***		0.3035 ***
Labour force status				
Employed (=reference)				
Not in labour force		0.1029		0.3028 ***
Not in labour force × Period		–0.2480		–0.3433 ***
Unemployed		0.3459 ***		1.2702 ***
Unemployed × Period		–0.2929		–1.3193 ***
Country of birth				
Born in Australia (=reference)				
Born overseas		0.0978		0.3254 ***
Adjusted R^2	0.1955	0.2008	0.2396	0.2599
RMSE	2.68	2.67	2.74	2.70

[+] Effect of 2007–08 relative to 1995.
* represents statistically significant at the 10% level of significance
** represents statistically significant at the 5% level of significance
*** represents statistically significant at the 1% level of significance

Weight misreporting

Table 5.2 shows the fitted models for weight misreporting. The negative intercepts suggest that weight is generally underestimated. The under-reporting is lower in 2007–08 than in 1995. The negative coefficients for the self-assessed overweight/obese variable implies that people who assess themselves as being overweight under-report their weight by more on average, even after taking account of their specific reported weight. The adjusted R^2 values, and thus the goodness-of-fit, for the weight models are even lower than for the height models. The adjusted R^2 values range from 3% to 6%, with females having a slightly higher model fit.

5.2 OLS model of weight misreporting, using pooled 1995 and 2007–08 data

	Male		*Female*	
	Simple	*Extended*	*Simple*	*Extended*
Dependent variable: Reported minus measured weight, in kg				
Intercept	–1.4091 ***	–1.4129 ***	–2.5177 ***	–2.5024 ***
Period	0.6745 ***	0.5614 ***	1.1692 ***	0.9832 ***
Reported weight	0.0254 ***	0.0297 ***	–0.0040	–0.0025
Reported height	–0.0288 ***	–0.0328 ***	–0.0236 ***	–0.0237 ***
Age	–0.0194 ***	–0.0201 ***	–0.0097 ***	–0.0115 ***
Age^2	0.0005 ***	0.0004 ***	0.0003 ***	0.0002 **
Self-assessed weight				
Underweight		1.4785 ***		1.0110 ***
Underweight × Period		–0.8419 ***		–0.1435
Acceptable (=reference)				
Overweight	–1.2672 ***	–1.5496 ***	–0.7288 ***	–0.9918 ***
Overweight × Period		0.5983 ***		0.6184 ***
Labour force status				
Employed or Not in labour force (=reference)				
Unemployed		0.3622 **		0.2069
Unemployed × Period		–0.0718		–1.0371 ***
Country of birth				
Born in Australia (=reference)				
Born overseas		–0.0659		0.1949 ***
Adjusted R^2	0.0348	0.0419	0.0532	0.0593
RMSE	3.57	3.55	3.08	3.07

BMI misreporting

Table 5.3 shows the fitted models for BMI misreporting. Females have greater reporting errors than males at the base case. Average under-reporting is again lower in 2007–08 than in 1995. The adjusted R^2 is higher for females than that for males, however, it is still low and only between 12% and 16% of the variation in misreporting of BMI is explained.

5.3 OLS model of BMI misreporting, using pooled 1995 and 2007–08 data

	Male		*Female*	
	Simple	*Extended*	*Simple*	*Extended*
Dependent variable: Reported minus measured BMI				
Intercept	–1.0696 ***	–1.0254 ***	–1.8430 ***	–1.6423 ***
Period	0.5424 ***	0.4764 ***	0.6431 ***	0.4944 ***
Reported BMI	0.3122 ***	0.3107 ***	0.3049 ***	0.3143 ***
Reported height	0.0472 ***	0.0445 ***	0.0379 ***	0.0391 ***
Reported weight	–0.0826 ***	–0.0799 ***	–0.1099 ***	–0.1125 ***
Age	–0.0214 ***	–0.0202 ***	–0.0231 ***	–0.0188 ***
Self-assessed weight				
Underweight		0.6390 ***		0.4749 ***
Underweight × Period		–0.3416 ***		–0.1297
Acceptable (=reference)				
Overweight	–0.5524 ***	–0.6407 ***	–0.4508 ***	–0.5251 ***
Overweight × Period		0.1919 ***		0.1626 ***
Self-assessed health				
Excellent, very good or good (=reference)				
Fair / poor health		–0.1320 ***		–0.1615 ***
Smoking status				
Current smoker (=reference)				
Never smoked		–0.0312		–0.1290 ***
Labour force status				
Employed (=reference)				
Not in labour force		–0.0283		–0.0730 **
Unemployed		–0.0367		–0.4333 ***
Adjusted R^2	0.1232	0.1300	0.1479	0.1574
RMSE	1.41	1.40	1.49	1.48

5.1.2 Comments on simple vs extended OLS

In reference to the tables in Section 5.1.1, the addition of socio-economic and other risk variables did not alter the signs and statistical significance of most of the original variables. Some of the additional variables were statistically significant for one of the sexes only. For example, country of birth was statistically significant for females but not significant for males in the regressions of height and weight. Other additional variables, such as self-assessed health status, were consistent across sexes in terms of statistical significance.

The adjusted R^2 values showed similar patterns for the Extended and Simple OLS models, with the height models having the highest values and the weight models having the lowest. The models for females again had higher R^2 values than the models for males. The Extended OLS models have marginally higher adjusted R^2 than the Simple OLS models, but they are still quite low for predictive models.

Appendix A shows the results of the models with age as a categorical variable. Replacing the continuous age variables with age groups does not change the signs and statistical significance of the majority of the variables. In fact, the coefficients and *p*-values were not materially affected. A number of age groups were found to be statistically insignificant in the regressions, especially for females in the reporting errors of height and weight, in contrast to the high significance when age is treated as a continuous variable. Further, the adjusted R^2 values from the regressions with age groups do not differ much from the adjusted R^2 values from the regressions with continuous age. As a result, there is little evidence that treating age as a categorical variable improves the predictive power of the models based on the analysis here.

The model based analysis done here confirms the results from the previous section that reporting errors were smaller in 2007–08 than in 1995. This is demonstrated by the coefficients for the Period variable (which indicates period effect on misreporting, i.e. difference between 2007–08 and 1995) having opposite signs to the intercept term in all the models (on the pooled dataset) and being statistically significant. The models also suggest that females have greater reporting errors than males. Even though the overall level of misreporting has declined, the other results from the individual datasets and the pooled dataset are consistent, suggesting the inferences from this section are reasonably robust.

However, the explanatory power of the models is low for predictive purposes. The highest adjusted R^2 observed was only 26%, with the models for weight having particularly low R^2 statistics.

The extended models had only slightly higher adjusted R^2 values than the simple models. Also, in comparison with the Simple OLS models, the Extended OLS did not alter the signs and statistical significance of most of the original variables from the Simple OLS models. At this stage, it does not seem like that the Extended OLS models would provide much benefit, compared to the Simple OLS models, for correcting for misreporting, despite using additional variables, which may not be available on other datasets that do not have measured BMI figures.

5.2 Semi-parametric regressions

Semi-parametric regression modelling is an alternative technique to linear regression modelling. It allows the slope of the relationship between the response variable and chosen explanatory variables to differ according to the values of those explanatory variables. The results presented in Section 4.5 showed that the direction of the relationship between the reporting error and the reported value of the measure of interest may differ according to the reported value of the measure (i.e. the relationship is not monotonic), implying that a linear regression may not fully capture this relationship. As such, a non-parametric term, which models the relationship in a more flexible manner, via an appropriate model, might provide a better fit to the data.

The type of semi-parametric model used in this section is a Generalised Additive Model (GAM), using a smoother fit to the reported height or reported weight, or both, depending on the response variable. The equations for the reporting error of each of the attribute are outlined as follows:

Height: $$h_{\text{rep},i} - h_{\text{act},i} = s\left(h_{\text{rep},i}\right) + s\left(w_{\text{rep},i}\right) + \sum_k \beta_k X_{ki} + \varepsilon_i$$

Weight: $$w_{\text{rep},i} - w_{\text{act},i} = s\left(w_{\text{rep},i}\right) + s\left(h_{\text{rep},i}\right) + \sum_k \beta_k X_{ki} + \varepsilon_i$$

BMI: $$b_{\text{rep},i} - b_{\text{act},i} = s\left(h_{\text{rep},i}, w_{\text{rep},i}\right) + \sum_k \beta_k X_{ki} + \varepsilon_i$$

where $s(h_i)$, $s(w_i)$ and $s(w_i, h_i)$ denote the smoothing terms, and the remaining variables, X_{ki}, enter parametrically with coefficients, β_k, to be estimated. The errors, ε_i, are assumed to be independently and identically distributed normal variables.

The GAMs estimated here use the same sets of variables as the Extended OLS models in Section 5.1.2, but replacing the reported attributes with the smoothers. Variables that were found to be not statistically significant in the model were removed. The coefficients are interpreted in the same way as the linear models. Note the coefficients presented here do not include height and weight, as these are included as non-parametric terms in the model (this means that the effect of height and weight on misreporting changes in a non-linear fashion over the range of height and weight in the data).

Tables 5.4–5.6 show the estimated coefficients for the models fitted to the pooled data[11], while Appendix B contains plots showing the estimated relationship between reporting errors and the reported value of the relevant attribute from these models. The results of the semi-parametric regressions for misreporting of weight and BMI are generally consistent with the Extended OLS regressions in terms of the values, signs and statistical significance of the coefficients for the regressions. However, for the regressions for misreporting of height, there were some notable differences in the coefficients for self-assessed body weight, with the effect of self-assessed overweight being smaller in the semi-parametric models and the effect of self-assessed underweight being larger in the semi-parametric models compared to the OLS models.

5.4 GAM model of height reporting error, using pooled 1995 and 2007–08 data

	Male	*Female*
Dependent variable: Reported minus measured height, in cm		
Intercept	1.6981 ***	0.1722 **
Period[+]	–1.0546 ***	–0.3708 ***
Age	0.0378 ***	0.0339 ***
Age^2	0.0011 ***	0.0015 ***
Self-assessed weight		
Underweight	0.0964	0.3292 **
Acceptable (=reference)		
Overweight	–0.2227 ***	0.1500 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	0.1400 *	0.3483 ***
Smoking status		
Current smoker (=reference)		
Never smoked	–0.1572 ***	0.3021 ***
Labour force status		
Employed (=reference)		
Not in labour force	0.1517	0.2650 ***
Not in labour force × Period	–0.2955 *	–0.3209 **
Unemployed	0.3602 ***	1.2486 ***
Unemployed × Period	–0.3083	–1.2561 ***
Country of birth		
Born in Australia (=reference)		
Born overseas	0.1365 **	0.3514 ***
Adjusted R^2	0.1910	0.2600
RMSE	2.6409	2.6722

[11] Regressions based on separate year data were also carried out, but the coefficients of the majority of the variables are consistent with the ones from the pooled dataset.

5.5 GAM model of weight reporting error, using pooled 1995 and 2007–08 data

	Male	*Female*
Dependent variable: Reported minus measured weight, in kg		
Intercept	–1.5104 ***	–2.3885 ***
Period	0.5631 ***	0.9886 ***
Age	–0.0177 ***	–0.0099 ***
Age^2	0.0003 **	0.0002 **
Self-assessed weight		
Underweight	1.5456 ***	1.1955 ***
Underweight × Period	–0.8178 **	–0.1112
Acceptable (=reference)		
Overweight	–1.4228 ***	–0.8968 ***
Overweight × Period	0.7133 ***	0.6637 ***
Labour force status		
Employed or Not in labour force (=reference)		
Unemployed	0.3816 ***	0.2102 *
Unemployed × Period	–0.0305	–1.0189 ***
Country of birth		
Born in Australia (=reference)		
Born overseas	–0.0087	0.2299 ***
Adjusted R^2	0.0417	0.0588
RMSE	3.5472	3.0606

The adjusted R^2 values are higher for females than males. However, none of the adjusted R^2 values for the semi-parametric models are substantially higher than the corresponding adjusted R^2 from the Extended OLS models, and many of them are slightly lower. The highest adjusted R^2 value among any of the semi-parametric models is only 26%, and they are particularly low for the weight models, with the highest being 6%. The adjusted R^2 for the semi-parametric models for BMI are marginally higher than those for the Extended OLS models.

The semi-parametric models were also fitted using age as a categorical variable and the results are in Appendix C. Similar to OLS, replacing continuous age variable with categorical age variable did not change the results materially. The signs and statistical significance of most of the variables and their *p*-values did not change considerably, and the adjusted R^2 statistics did not differ substantially. With regards to the coefficients for the different age groups, a number of age groups were not statistically significant, especially in the weight model.

5.6 GAM model of BMI reporting error, using pooled 1995 and 2007–08 data

	Male	*Female*
Dependent variable: Reported minus measured BMI		
Intercept	–0.9214 ***	–0.9349 ***
Period	0.4561 ***	0.4715 ***
Age	–0.0204 ***	–0.0196 ***
Self-assessed weight		
Underweight	0.8004 ***	0.6707 ***
Underweight × Period	–0.3459 ***	–0.1431
Acceptable (=reference)		
Overweight	–0.6510 ***	–0.5691 ***
Overweight × Period	0.2598 ***	0.2115 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	–0.1136 ***	–0.1522 ***
Smoking status		
Current smoker (=reference)		
Never smoked	–0.0315	–0.1263 ***
Labour force status		
Employed (=reference)		
Not in labour force	–0.0195	–0.0640 *
Unemployed	–0.0299	–0.4239 ***
Adjusted R^2	0.1440	0.164
RMSE	1.3892	1.4734

In conclusion, the results from the GAM models for misreporting of weight and BMI were generally consistent with the OLS models, but for some self-assessed body weight and health variables they were found to be different from the OLS models for the regressions of misreporting of height.

Given that neither has a clear-cut advantage over the other, it is difficult to choose among these two alternative sets of models. The simplicity and ease of interpretation of the OLS models has to be balanced against the sophistication of the semi-parametric model. The two methods are assessed further in Section 6 by seeing how well they estimate the proportion in each BMI category.

6. ADJUSTING BMI CATEGORY ESTIMATES

In this section we apply the models examined in Section 5 to calculate corrected BMI distributions from self-reported BMI, using NHS 2007–08 data, and compare these results to the measured and reported estimates. First, the models are applied to the sample of records that have both measured and reported figures available, and the adjusted estimates are compared to corresponding measured and reported estimates to assess the suitability of the various models. We then calculate adjusted BMI for those records that do not have measured data (but do have reported data) by imputing values using the models and combine it with those records that have the measured data to calculate the adjusted BMI for the total (18+) population.

The estimated regression models are used to adjust the self-reported values to obtain estimates of each attribute as follows:

$$y_{\text{adj},i} = y_{\text{rep},i} - \left(\hat{d}_i + \hat{\sigma} \cdot e_i\right)$$

where:

$y_{\text{adj},i}$ is the adjusted estimate of the relevant attribute for record i;

$y_{\text{rep},i}$ is the self-reported estimate of the relevant attribute for record i;

$\hat{d}_i$ is the value of d_i predicted from the regression;

d_i $= y_{\text{rep},i} - y_{\text{act},i}$;

$y_{\text{act},i}$ is the measured value of the relevant attribute for record i;

$\hat{\sigma}$ is the estimated root mean squared error (RMSE) of the regression;

e_i is a random term, drawn from a standard normal distribution.

The addition of the random term, e_i, accounts for the variability of d_i that is not captured by the model. Omitting the random component would result in the variability of the measured values being underestimated, unless the model predicts d_i perfectly.

6.1 Assessing the accuracy of corrected estimates

Here we assess the accuracy of the various models by comparing the adjusted estimates with the estimates from measured and reported values using data from the sample of records (aged 18+) from NHS 2007–08 that have both measured and reported figures available.[12] As part of this assessment we examine which methods provide better estimates of BMI: smaller or larger models (Simple OLS *vs* Extended OLS); alternative modelling techniques (OLS *vs* Semi-parametric); alternative

[12] Here we add back the outliers that were removed when we estimated the models.

specifications of the age variable (continuous *vs* categorical); and alternative methods of adjusting BMI (directly or indirectly from adjusted height and weight).[13]

Table 6.1 specifies the models used to compute the adjusted BMI category proportions.

6.1 Description of models computing BMI category

Measured	BMI derived from measured height and weight
Reported	BMI derived from reported height and weight
S_OLS (cont. age)	Derive adjusted BMI directly from reported BMI using Simple OLS with Age as a continuous variable
S_OLS (h w)	Derive adjusted height and weight using Simple OLS with Age as a continuous variable, then derive BMI from adjusted figures
OLS (cont. age)	Derive adjusted BMI directly from reported BMI using Extended OLS with Age as a continuous variable
OLS (h w)	Derive adjusted height and weight using Extended OLS with Age as a continuous variable, then derive BMI from adjusted figures
OLS (age group)	Derive adjusted BMI directly from reported BMI using Extended OLS with Age as a categorical variable
OLS (h w age group)	Derive adjusted height and weight using Extended OLS with Age as a categorical variable, then derive BMI from adjusted figures
GAM (cont. age)	Derive adjusted BMI directly from reported BMI using GAM with Age as a continuous variable
GAM (h w)	Derive adjusted height and weight using GAM with Age as a continuous variable, then derive BMI from adjusted figures
GAM (age group)	Derive adjusted BMI directly from reported BMI using GAM with Age as a categorical variable
GAM (h w age group)	Derive adjusted height and weight using GAM with Age as a categorical variable, then derive BMI from adjusted figures

The same random values, e_i, were used to generate the adjusted height values for each of the height models. The same was done for each of the weight models and each of the BMI models.[14] This was done to minimise the amount of difference between the estimates from different methods that is due to random noise.

Table 6.2 presents the weighted BMI distributions for the sample of records that have both measured and reported figures, calculated from the measured, reported and adjusted BMI figures respectively. Table 6.3 shows the ratio of adjusted to measured estimates of each BMI category for each of the models. By using only records that provided both measured and reported height and weight, we can assess how close the different correction models get to the results using measured data, without non-response bias affecting the comparisons.

13 Note as a further assessment of the robustness of the models we also undertook some further analysis where we estimated the Extended OLS models using a sub-sample of the data (randomly selected 50% sample) and applied to the remainder of the sample to compute the BMI categories. The regression results from the sub-sample were consistent with the results from the full sample and adjusted BMI distributions were close to those presented in this paper.

14 Three different sets of e_i values were used: one for the height models; one for the weight models; and one for the BMI models.

6.2 Adjusted BMI category estimates 2007–08, sample with both reported and measured BMI

	Underweight	*Normal*	*Overweight*	*Obese*	*Total*
Measured	0.0199	0.3718	0.3673	0.2410	1.0000
Reported	0.0261	0.4243	0.3452	0.2044	1.0000
S_OLS (cont. age)	0.0243	0.3639	0.3633	0.2485	1.0000
S_OLS (h w)	0.0243	0.3712	0.3561	0.2484	1.0000
OLS (cont. age)	0.0265	0.3571	0.3678	0.2486	1.0000
OLS (h w)	0.0244	0.3652	0.3605	0.2498	1.0000
OLS (age group)	0.0269	0.3654	0.3593	0.2484	1.0000
OLS (h w age group)	0.0243	0.3610	0.3672	0.2475	1.0000
GAM (cont. age)	0.0216	0.3680	0.3657	0.2447	1.0000
GAM (h w)	0.0208	0.3617	0.3683	0.2493	1.0000
GAM (age group)	0.0257	0.3631	0.3635	0.2478	1.0000
GAM (h w age group)	0.0237	0.3691	0.3595	0.2477	1.0000

6.3 Ratio of adjusted to measured BMI category estimates 2007–08, sample with both reported and measured BMI

	Underweight	*Normal*	*Overweight*	*Obese*
Reported	1.3104	1.1413	0.9398	0.8480
S_OLS (cont. age)	1.2207	0.9788	0.9891	1.0312
S_OLS (h w)	1.2207	0.9984	0.9695	1.0308
OLS (cont. age)	1.3313	0.9605	1.0013	1.0316
OLS (h w)	1.2258	0.9823	0.9814	1.0366
OLS (age group)	1.3513	0.9828	0.9782	1.0308
OLS (h w age group)	1.2207	0.9710	0.9997	1.0271
GAM (cont. age)	1.0851	0.9898	0.9956	1.0154
GAM (h w)	1.0449	0.9729	1.0027	1.0345
GAM (age group)	1.2911	0.9766	0.9896	1.0283
GAM (h w age group)	1.1906	0.9928	0.9787	1.0279

All correction models provide estimates that are closer to the estimates from the measured BMI than those from the self-reported figures. Most of the correction models are similar in terms of overall accuracy. However, the GAM models estimating BMI directly, with age as a continuous variable, stands out as the most accurate, especially for the Obese and Overweight categories, which are of primary interest.

There are some differences in the estimates between using age as a continuous variable compared to using it as a categorical variable, but neither is consistently better than the other. Little improvement, if any, was seen between the estimates that used BMI obtained indirectly from adjusted height and weight to those obtained directly, but this could to some extent reflect the poor fit for the weight equation.

Of some concern is the fact that all the correction models overestimated the proportions in the Underweight and Obese categories, and slightly underestimated the proportion in the Normal category.

6.2 BMI category estimates for the population

Here the BMI distribution for the population aged 18 and over is estimated from the NHS 2007–08, using measured BMI for those records that have it and imputed BMI for those records that have self-reported BMI but no measured BMI. These estimates attempt to correct for both the reporting bias in self-reported figures and for any selection bias in the measured data.

The adjusted estimates for the GAM models with age as a continuous variable, estimating BMI directly, along with the distributions estimated using measured and self-reported figures, are shown in table 6.4. The adjusted estimates using the other models are presented in Appendix D, but the results do not vary much between the different models. The proportion estimated to be Obese or Thin is higher, and the proportion estimated to be Overweight or Normal is lower, in the adjusted distribution than in the distribution of measured BMI. This is the same pattern of discrepancy as in tables 6.2 and 6.3, so it is hard to say how much of the difference is due to problems in the correction models, and how much is due to some groups being under-represented in the measured data.

6.4 Adjusted BMI category estimates 2007–08

	Measured	*Reported*	*GAM (cont. age)*
Thin	0.0202	0.0249	0.0203
Normal	0.3686	0.4174	0.3651
Overweight	0.3664	0.3446	0.3595
Obese	0.2448	0.2132	0.2550
Total	1.0000	1.0000	1.0000

7. CONCLUSION AND RECOMMENDATIONS

The results presented in this paper show that, in accordance with previous studies, individuals tend to over-report their height and under-report their weight. These effects lead to under-reporting of BMI and an underestimation of the prevalence of obesity in the population when self-reported figures are used.

Reporting errors for both males and females were found to be smaller in 2007–08 than in 1995. However, as data are available only for two time points, there is not enough information to say that there is a general decreasing trend over time.

The analysis shows evidence that females tend to have greater reporting errors than males. The size of the reporting errors were also found to depend on the reported values of height, weight and BMI, as well as various demographic and socio-economic variables, and other health-related risk factors. The amount of the variation in reporting error that could be explained by the models examined was only moderate in the case of the reporting error for height, and very low in the case of the reporting error for weight. However, replacing reporting error with measured values largely improved the R^2 (regressions of measured value on the corresponding reported value yield R^2 of at least 80%). This suggests that reported value is a good predictor of measured value and the low R^2 of our models should not be a concern.

Given the observed reporting bias, self-reported BMI figures should be adjusted to get better estimates of the true BMI distribution. The estimated regression models were used to impute measured BMI for each individual based on their reported values and other characteristics, rather than applying the same correction for all persons. Adjusted estimates of the BMI distribution were then calculated using the imputed BMI values.

The accuracy of the adjusted estimates was assessed by estimating the proportion of persons in each BMI category using only the records that had both measured and self-reported BMI. These adjusted estimates were then compared to the corresponding estimates from measured and reported data.

The corrected BMI category proportions from all correction models were found to be closer to the measured BMI category proportions than those using the original reported BMI. While all correction models examined here overestimated the proportions that were obese or underweight, and underestimated the proportions that were overweight or normal, the discrepancies between the adjusted and measured values were much lower than that between the reported and measured values. Deriving BMI from adjusted height and weight, rather than adjusting the reported BMI directly, made only a small, if any, improvement although this could reflect the poor fit for the weight equation.

The semi-parametric models overall did not perform greatly better than the OLS models although the correction model identified as giving the most accurate estimates was a semi-parametric model. Given the additional complexity of the semi-parametric models, and that the improvement in the accuracy of the BMI category estimates was only minor, the linear regression models might be preferred to the semi-parametric models for adjusting BMI estimates.

While the results in this paper suggest that adjusting self-reported BMI data can improve its accuracy, the adjustments have only been applied to the same data that was used to fit the correction models. An outstanding question is whether the models fitted to this data could be used to adjust BMI for future surveys that may not collect any measured height and weight data. While a substantial difference in the level of misreporting was found between 1995 and 2007–08, these surveys were a relatively long period apart, and most of the coefficients in the regressions were quite similar between the two periods. More data may be necessary to assess whether the levels and pattern of misreporting have stabilised before extrapolating these models to future surveys. However, future data may prove that misreporting has not stabilised, and hence, it would be very risky to extrapolate the models to future surveys at this stage. With regards to model choice and specification a range of models were examined in this paper and they all gave similar results in terms of BMI distribution. As such it is unlikely that the estimates can be improved by further research in modelling method, but we do not preclude the possibility to improve the estimates with extra explanatory variables.

The results in this paper demonstrate that it is necessary to adjust self-reported BMI data for reporting biases. They also suggest that, at least for the NHS 2007–08, supplementing measured data with adjusted self-reported data may improve estimates of obesity prevalence in the population. However, some discrepancies between the adjusted and measured estimates were found when comparing them using the same sample and that care still needs to be taken when using and interpreting the adjusted data.

REFERENCES

Australian Bureau of Statistics (1995) *How Australians Measure Up,* cat. no. 4359.0, ABS, Canberra.

Australian Bureau of Statistics (1995, 2007–08) *National Health Survey, Australia,* cat. no. 4364.0, ABS, Canberra.

Dauphinot, V.; Wolff, H.; Naudin, F.; Gueguen, R.; Sermet, C.; Gaspoz, J–M. and Kossovsky, M.P. (2009) "New Obesity Body Mass Index Threshold for Self-reported Data", *Journal of Epidemiology and Community Health*, 63, pp. 128–132.

Hayes, A.J.; Kortt, M.A.; Clarke, P.M. and Brandrup, J.D. (2008) "Estimating Equations to Correct Self-reported Height and Weight: Implications for Prevalence of Overweight and Obesity in Australia", *Australian and New Zealand Journal of Public Health*, 32(6), pp. 542–545.

McAdams, M.A.; Van Dam, R.M. and Hu, F.B. (2007) "Comparison of Self-reported and Measured BMI as Correlates of Disease Markers in U.S. Adults", *Obesity*, 15(1), pp. 188–196.

Wang, Z.; Patterson, C.M. and Hills, A.P. (2002) "A Comparison of Self-Reported and Measured Height, Weight and BMI in Australian Adolescents", *Australian and New Zealand Journal of Public Health*, 26(5), pp. 473–478.

ACKNOWLEDGEMENTS

The authors would like to thank Associate Professor Robert Clark of the Centre for Statistical and Survey Methodology, University of Wollongong and Ms. Louise Gates, Director of Health and Disability Section, Australian Bureau of Statistics for their advice and comments. Responsibility for any errors or omissions remains solely with the authors.

APPENDIXES

A. EXTENDED MODELS WITH AGE GROUPS, USING POOLED 1995 AND 2007–08 DATA

A.1 Extended OLS model of height misreporting with age groups, using pooled 1995 and 2007–08 data

	Male	*Female*
Intercept	0.6262 ***	0.9563 ***
Period	–0.9040 ***	–0.2518 ***
Reported height	0.1944 ***	0.1877 ***
(Reported height)^2	–0.0036 ***	0.0000
Reported weight	–0.0403 ***	–0.0277 ***
Age		
18–24 years	–0.3648 ***	–0.0831
25–34 years	–0.1117	–0.0934
35–44 years (=reference)		
45–54 years	0.2917 ***	0.2186 **
55–64 years	0.9176 ***	0.9190 ***
65–74 years	1.6372 ***	1.5711 ***
75+ years	2.8421 ***	3.1834 ***
Self-assessed weight		
Underweight	–0.3553	0.0943
Acceptable (=reference)		
Overweight	0.3667	0.5828 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	0.2678	0.4364 ***
Smoking status		
Current smoker (=reference)		
Never smoked	–0.1788	0.3152 ***
Labour force status		
Employed (=reference)		
Not in labour force	0.1842	0.3337 ***
Not in labour force × Period	–0.3058 *	–0.3007 **
Unemployed	0.3716	1.3621 ***
Unemployed × Period	–0.3359	–1.3987 ***
Country of birth		
Born in Australia		
Born overseas	0.1009	0.3264 ***
Adjusted R^2	0.1986	0.2571
RMSE	2.6701	2.7049

A.2 Extended OLS model of weight misreporting with age groups, using pooled 1995 and 2007–08 data

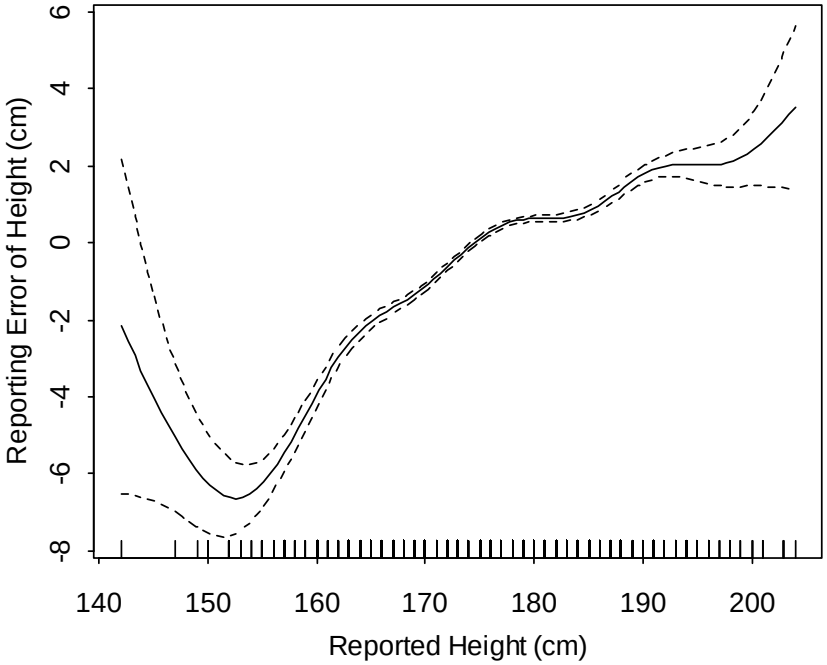
	Male	*Female*
Intercept	–1.1670 ***	–2.4995 ***
Period	0.5356 ***	1.0102 ***
Reported weight	0.0297 ***	–0.0025
Reported height	–0.0322 ***	–0.0238 ***
Age		
18–24 years	0.4161 ***	0.3799 ***
25–34 years	0.1653	0.2581 ***
35–44 years (=reference)		
45–54 years	–0.4937 ***	–0.1157
55–64 years	–0.4583 ***	0.0305
65–74 years	–0.3394 **	–0.3379 **
75+ years	–0.3796 **	–0.1830
Self-assessed weight		
Underweight	1.4824 ***	1.0081 ***
Underweight × Period	–0.8268 **	–0.1384
Acceptable (=reference)		
Overweight	–1.5499 ***	–0.9934 ***
Overweight × Period	0.5950 ***	0.6156 ***
Labour force status		
Employed or Not in labour force (=reference)		
Unemployed	0.2358	0.3325 **
Unemployed × Period	0.0827	–1.1741 ***
Country of birth		
Born in Australia		
Born overseas	–0.0669	0.1942 ***
Adjusted R^2	0.0420	0.0598
RMSE	3.5547	3.0644

A.3 Extended OLS Model of BMI misreporting with age groups, using pooled 1995 and 2007–08 data

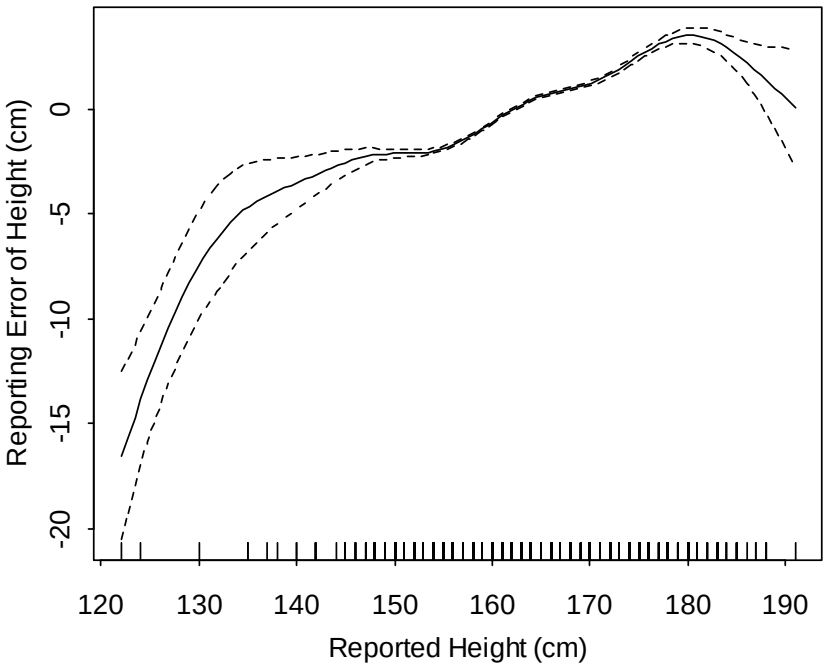
	Male	*Female*
Intercept	–0.8609 ***	–1.5368 ***
Period	0.4809 ***	0.5229 ***
Reported BMI	0.3145 ***	0.3139 ***
Reported height	0.0464 ***	0.0400 ***
Reported weight	–0.0814 ***	–0.1134 ***
Age		
18–24 years	0.2485 ***	0.1708 ***
25–34 years	0.0937 **	0.1255 ***
35–44 years (=reference)		
45–54 years	–0.2563 ***	–0.1306 ***
55–64 years	–0.4645 ***	–0.3324 ***
65–74 years	–0.6344 ***	–0.6781 ***
75+ years	–0.9720 ***	–1.0966 ***
Self-assessed weight		
Underweight	0.6492 ***	0.4720 ***
Underweight × Period	–0.3431 ***	–0.1317
Acceptable		
Overweight	–0.6479 ***	–0.5397 ***
Overweight × Period	0.1885 ***	0.1687 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	–0.1350 ***	–0.1503 ***
Smoking status		
Current smoker		
Never smoked	–0.0227	–0.1149 ***
Labour force status		
Employed (=reference)		
Not in labour force	–0.0193	–0.0380
Unemployed	–0.0235	–0.2852 ***
Adjusted R^2	0.1279	0.1601
RMSE	1.4016	1.4784

B. ESTIMATED SMOOTH TERMS FROM GAM MODELS

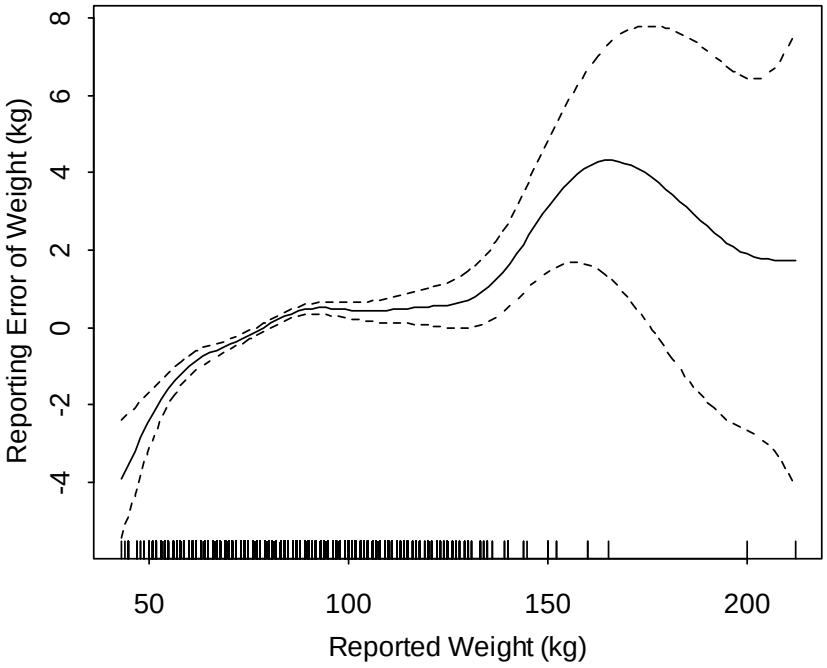
B.1 Expected Height Reporting Error *vs* Reported Height, Males, Pooled



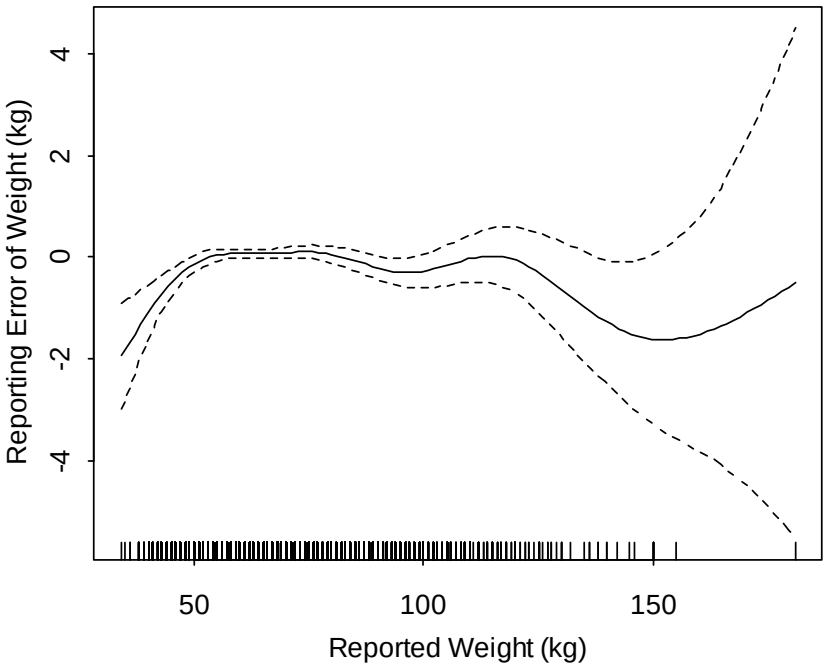
B.2 Expected Height Reporting Error *vs* Reported Height, Females, Pooled



B.3 Expected Weight Reporting Error *vs* Reported Weight, Males, Pooled



B.4 Expected Weight Reporting Error *vs* Reported Weight, Females, Pooled



C. GAM MODELS WITH AGE GROUPS

C.1 GAM model of height reporting error with age groups, using pooled 1995 and 2007–08 data

	Male	*Female*
Intercept	1.3588 ***	–0.1737 **
Period	–0.9378 ***	–0.2565 ***
Age		
18–24 years	–0.4055 ***	–0.1370
25–34 years	–0.1089	–0.1283
35–44 years (=reference)		
45–54 years	0.3080 ***	0.2615 ***
55–64 years	0.9452 ***	0.9902 ***
65–74 years	1.7173 ***	1.6816 ***
75+ years	2.8692 ***	3.2297 ***
Self-assessed weight		
Underweight	–0.7004 ***	–0.3314 **
Acceptable (=reference)		
Overweight	0.3980 ***	0.6881 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	0.2242 ***	0.4041 ***
Smoking status		
Current smoker (=reference)		
Never smoked	–0.1788 ***	0.3113 ***
Labour force status		
Employed (=reference)		
Not in labour force	0.1348	0.3214 ***
Not in labour force × Period	–0.2949 *	–0.3173 **
Unemployed	0.2939 **	1.3084 ***
Unemployed × Period	–0.3581	–1.2825 ***
Country of birth		
Born in Australia (=reference)		
Born overseas	0.0830	0.2959 ***
Adjusted R^2	0.214	0.272
RMSE	2.6445	2.6784

C.2 GAM model of weight reporting error with age groups, using pooled 1995 and 2007–08 data

	Male	*Female*
Intercept	–1.1907 ***	–2.3024 ***
Period	0.5171 ***	0.9925 ***
Age		
18–24 years	0.4616 ***	0.4216 ***
25–34 years	0.1691	0.2727 ***
35–44 years (=reference)		
45–54 years	–0.4941 ***	–0.1316
55–64 years	–0.4694 ***	0.0083
65–74 years	–0.3717 **	–0.3609 ***
75+ years	–0.3724 **	–0.1834
Self-assessed weight		
Underweight	1.7548 ***	1.3046 ***
Underweight × Period	–0.8377 ***	–0.1044
Acceptable (=reference)		
Overweight	–1.5933 ***	–1.0477 ***
Overweight × Period	0.6965 ***	0.6741 ***
Labour force status		
Employed or Not in labour force (=reference)		
Unemployed	0.2624 *	0.3429 **
Unemployed × Period	0.0924	–1.1789 ***
Country of birth		
Born in Australia (=reference)		
Born overseas	–0.0536	0.2100 ***
Adjusted R^2	0.046	0.0625
RMSE	3.5472	3.0600

C.3 GAM model of BMI reporting error with age groups, using pooled 1995 and 2007–08 data

	Male	*Female*
Intercept	–0.7513 ***	–0.8377 ***
Period	0.4620 ***	0.5101 ***
Age		
18–24 years	0.2689 ***	0.2039 ***
25–34 years	0.0914 **	0.1330 ***
35–44 years (=reference)		
45–54 years	–0.2531 ***	–0.1404 ***
55–64 years	–0.4594 ***	–0.3434 ***
65–74 years	–0.6464 ***	–0.6892 ***
75+ years	–0.9739 ***	–1.0935 ***
Self-assessed weight		
Underweight	0.8039 ***	0.7032 ***
Underweight × Period	–0.3516 ***	–0.1396
Acceptable (=reference)		
Overweight	–0.6572 ***	–0.5679 ***
Overweight × Period	0.2519 ***	0.2071 ***
Self-assessed health		
Excellent, very good or good (=reference)		
Fair / poor health	–0.1175 ***	–0.1417 ***
Smoking status		
Current smoker (=reference)		
Never smoked	–0.0237	–0.1137 ***
Labour force status		
Employed (=reference)		
Not in labour force	–0.0100	–0.0321
Unemployed	–0.0130	–0.2824 ***
Adjusted R^2	0.141	0.168
RMSE	1.3910	1.4718

D. ADJUSTED BMI CATEGORY ESTIMATES

D.1 Adjusted BMI category estimates, 2007–08

	Thin	*Normal*	*Overweight*	*Obese*	*Total*
S_OLS (cont. age)	0.0228	0.3590	0.3642	0.2540	1.0000
S_OLS (h w)	0.0205	0.3626	0.3609	0.2560	1.0000
OLS (cont. age)	0.0205	0.3628	0.3608	0.2559	1.0000
OLS (h w)	0.0193	0.3636	0.3615	0.2557	1.0000
OLS (age group)	0.0199	0.3609	0.3646	0.2546	1.0000
GAM (cont. age)	0.0203	0.3651	0.3595	0.2550	1.0000
GAM (h w)	0.0196	0.3616	0.3660	0.2528	1.0000
GAM (age group)	0.0200	0.3613	0.3638	0.2549	1.0000
GAM (h w age group)	0.0196	0.3639	0.3624	0.2541	1.0000

E. BMI CATEGORY CLASSIFICATION

E.1 BMI category classification

Thin	<18.5
Normal	18.5 – 25
Overweight	25 – 30
Obese	>30

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au