Estimation of Change in Learning Disability Statistics Scotland

Matthew Greenaway
Methodology Advisory Service, Office for National Statistics
Quality Improvement Fund Report, March 2016

1. Background

Learning Disability Statistics Scotland is an annual publication containing statistics on adults with learning disabilities known to Scottish local authorities. The publication is based on administrative records – every year LDSS request data from each Scottish local authority on all adults with learning disabilities held on that local authority's information management system. A number of characteristics are requested on each individual, including 'attribute data' such as age and sex and variables of interest such as autism spectrum diagnosis and accommodation type.

Local authorities provide at least some information on all adults with learning disability known to them (except for rare occasions when no response at all is received), but much of the data on variables of interest is incomplete – for example, of the 26,786 adults reported for in 2014, 4,048 had a missing autism spectrum diagnosis. This missingness is not equally spread between local authorities, with in some cases local authorities reporting very little or no data for a particular variable. This makes reporting change over time challenging, particularly as the overall amount of missingness has reduced over time.

Prior to the 2014 publication time-series were reported throughout the main publication accompanied by caveats but no adjustments for missingness. In many cases, the time series for totals showed a potentially misleading increase over time simply due to a reduction in missingness - an example of this is given in section 2. As a condition of the publication receiving National Statistics accreditation, time series analysis was removed entirely from the 2014 report.

This document summarises options for estimation and reporting of change for future publications of Learning Disability Statistics Scotland, focusing on weighting as opposed to imputation to adjust for missingness. Analysis on missingness in the dataset is contained in section 3, some options for reporting of change are in section 4, and recommendations in section 5.

2. The challenge of reporting change - example

This section includes a brief example to illustrate the challenge of reporting change.

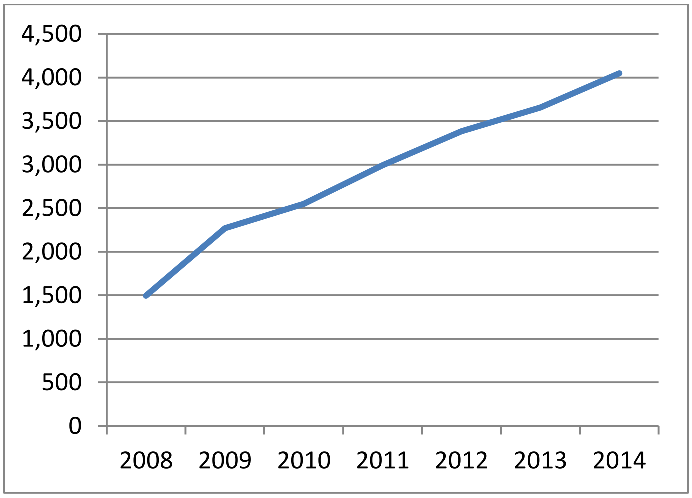
The table below contains the number of adults recorded with & without an autism spectrum diagnosis from 2008 to 2014, and the number of adults with unknown autism spectrum diagnosis. A similar table appeared in the 2013 report (without the 2014 figures), but no comparisons over time were included in the 2014 report.

Table 1: autism spectrum diagnosis over time

	Counts of adults known to local authorities		
	Have Autism Spectrum Diagnosis (A)	**Do not have Autism Spectrum Diagnosis (B)**	**Unknown (C)**
2008	1,494	11,957	11,801
2009	2,270	13,547	11,854
2010	2,548	17,656	7,187
2011	2,992	17,924	5,120
2012	3,385	18,291	4,441
2013	3,655	18,053	4,528
2014	4,048	18,260	4,478

A simple line graph of the total number of adults with autism spectrum diagnosis (column A in the table above) would misleadingly show the total increasing dramatically from 2008 to 2014. This is clearly due mostly to the fact that the ‘unknown’ category has reduced over time. More generally, estimating totals by simply counting non-missing data will systematically under-estimate true totals, and the degree of under-estimation will vary over time as missingness changes.

Graph 1: autism spectrum diagnosis total over time: totals



An alternative would be to report proportions instead of totals. Where this has been done in previous LDSS bulletins, the missing data in the denominator of the proportion (table 2), but an alternative would be to calculate proportions excluding missing data (table 3).

Table 2: including missing data in denominator, 2013 and 2014 only

	Proportion with Autism Spectrum Diagnosis (A/A+B+C)*100	**Proportion no Autism Spectrum Diagnosis (B/A+B+C)*100**	**Proportion Missing (C/A+B+C)*100**	**Total**
2013	13.93%	68.81%	17.26%	100%
2014	15.11%	68.17%	16.72%	100%

Table 3: excluding missing data, 2013 and 2014 only

	Proportion non-missing with Autism Spectrum Diagnosis	**Proportion non-missing no Autism Spectrum Diagnosis**	**Total**	*Number of missing responses*
2013	16.84%	83.16%	100%	*4,528*
2014	18.15%	81.85%	100%	*4,478*

The approach in table 3 – including only non-missing data in the calculation of the proportion - involves implicitly estimating for missing respondents using non-missing respondents, which may not be appropriate. However, table 2 suffers from many of the same issues as graph 1 – proportions are difficult to compare over time due to changing missingness.

Whether either of these methods are appropriate, or whether a weighting method which might allow unbiased estimation of change exists, depends on the drivers of missingness in the dataset. This is discussed in the next section.

3. Missingness in the dataset

Data may be missing –

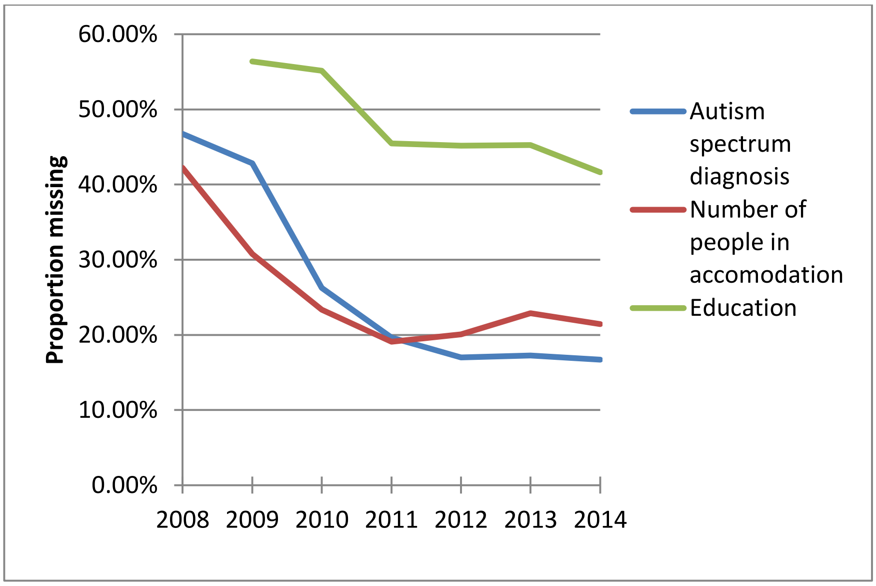
- 'completely at random' – meaning that missingness is simply random, and does not depend on any other observed or unobserved variables
- 'at random' – meaning that missingness is at random when controlling for observed variables – for example, attribute data like age and sex
- 'not at random' – meaning that missingness depends directly on the variables being measured.

Weighting a dataset will remove bias where the variables used in the weighting are correlated with both the outcome variables and the missingness mechanism. For example, if age is correlated with both learning disability outcomes and with the probability of a record being missing, then weighting using age will remove bias in learning disability estimates. Another way of putting this is that weighting is beneficial where the data is missing at random with respect to the variables used in the weighting, and outcome variables are correlated with weighting variables.

Long-term missingness trends

The chart below summarises the longer-term trend in missingness for three variables – autism spectrum diagnosis, number of people in accommodation, and education. The general pattern is a fairly sharp drop in missingness in the earlier years the survey was running – 2008-2011 – and of a levelling-off in recent years.

Graph 2: change in percentage missing since 2008



Missingness with respect to age and sex

A study on the potential use of imputation on the earlier 'ESAY' survey[1] investigated the relationship between the variables being measured and the 'attribute data' available for all cases – age, sex and ethnicity. The conclusion presented in the paper was that non-response is spread fairly evenly across these variables, implying that using these variables in either a weighting or imputation approach would not remove bias due to missingness. For example, the table below gives the percentage missing for a number of variables by gender, and there are no large differences apparent.

Table 4: missingness by gender

	Percent missing	
	Male	**Female**
Autism spectrum diagnosis	16.50%	17.80%
Person service	10.90%	10.90%
Employment status	33.10%	31.50%
Day care centre attendance	14.30%	13.50%
Accommodation Type	9.20%	8%

Utilising age or sex in a weighting approach is therefore unlikely to remove bias from estimates.

Missingness with respect to local authorities

The table on the next page gives the percentage completeness for each item for each local authority in 2012, 2013, and 2014. There are several patterns worth noting –

- Many local authorities have consistently high missingness for some variables and low missingness for others. For example, West Lothian and the Shetland Islands have consistently high missingness for autism spectrum diagnosis but consistently low missingness elsewhere.
- A smaller number of local authorities provide good response for a variable in some years but poorer response in others. For example, Clackmannanshire provided good-quality data for Employment Opportunities in 2012 and 2014, but not in 2013.
- A limited number of local authorities have uniformly high missingness across most variables – for example, the Highlands
- For East Renfrewshire in 2014, no data at all is available

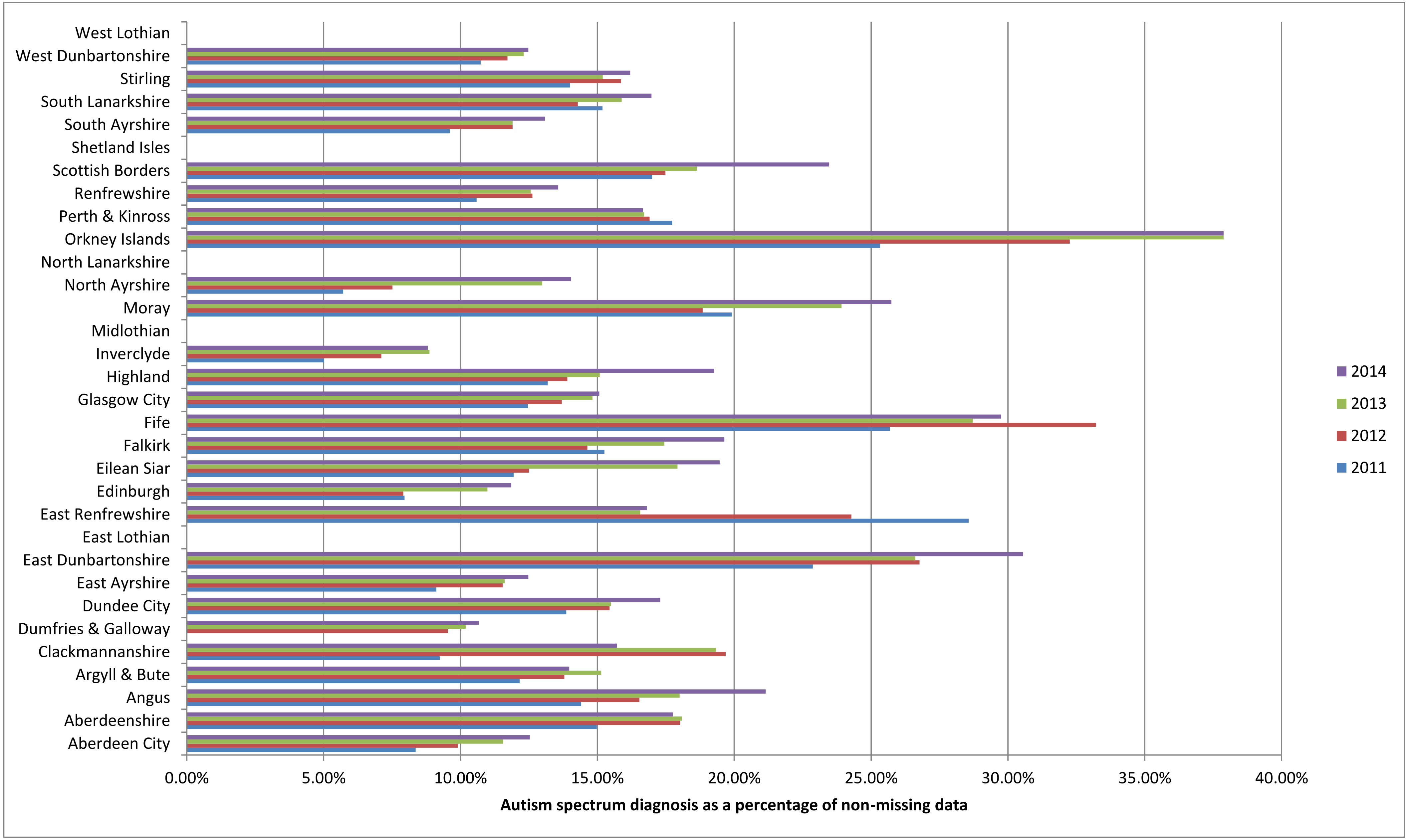
This suggests that a mix of factors may be driving missingness – some local authorities may simply not have data available, while a smaller number may vary in their reporting of the data they have year-on-year. It is, however, fairly clear that missingness varies by local authority.

Missingness varying by local authority will cause bias in estimates if learning disability outcomes vary by local authority. Graph 1 shows the variation in autism spectrum diagnosis by LA for 2011-2014. There is some evidence that this characteristic does vary by LA – variation in estimates between LAs is much larger than variation within LAs.

[1] Miltilado, M. and Wardman, L. "An Assessment of the potential for imputation of non-response in the eSAY survey", available at https://gss.civilservice.gov.uk/wp-content/uploads/2014/09/Final-report-on-feasibility-and-data-imputation-on-eSAY-dataset.pdf

Local authority	AS diagnosis			Family carer			Number in same household			Accommodation type			LAC			PLP			Employment opportunities			Day centre			Alternative opportunities			Further education		
	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14	'12	'13	'14
Aberdeen City	100	100	100	79	89	92	74	85	88	93	93	96	100	100	100	68	80	83	73	80	83	97	100	89	63	76	79	63	82	85
Aberdeenshire	81	78	74	55	22	28	52	0.4	0	81	30	33	100	100	100	66	64	62	38	19	19	27	24	100	14	0.3	5	14	0.4	0
Angus	94	94	97	91	92	97	84	86	89	95	95	99	100	100	100	90	88	95	90	91	96	90	91	96	90	89	93	90	94	96
Argyll & Bute	97	98	97	97	98	98	100	99	99	99	99	99	100	100	100	98	97	97	96	97	97	96	97	98	96	97	97	96	96	96
Clackmannanshire	100	100	100	96	92	0	0	0.4	0	100	100	100	100	100	100	100	100	100	100	9.7	100	100	100	100	27	100	100	27	1.1	100
Dumfries & Galloway	64	65	68	98	99	99	92	94	94	100	100	100	100	100	100	100	100	100	86	89	90	85	86	87	46	51	55	46	91	91
Dundee City	84	78	83	89	83	91	82	77	84	93	87	99	100	100	100	83	85	86	87	81	80	91	85	83	95	90	83	95	77	87
East Ayrshire	86	95	100	97	98	99	94	97	97	100	100	100	100	100	100	97	100	100	83	86	95	85	99	100	98	88	100	98	90	96
East Dunbartonshire	100	100	100	100	100	99	99	99	99	100	100	99	100	100	100	97	93	93	90	90	83	91	88	90	90	86	69	90	88	21
East Lothian	32	43	43	100	100	100	96	95	95	100	100	100	100	100	100	68	67	66	70	69	68	100	100	100	69	63	63	69	59	60
East Renfrewshire	98	99	0	99	99	0	98	99	0	99	99	0	100	100	0	94	96	0	99	98	0	98	95	0	98	94	0	98	93	0
Edinburgh	100	100	100	52	82	85	86	83	92	88	84	92	100	100	100	23	14	3	20	14	22	100	9.7	9	11	9.7	9	11	100	0
Eilean Siar	99	99	99	94	97	99	88	89	91	94	99	98	100	100	100	92	98	98	73	94	98	88	96	100	84	85	87	84	92	93
Falkirk	84	82	73	83	88	82	78	78	72	83	90	82	100	100	100	71	57	58	76	79	74	73	77	73	76	74	68	76	84	80
Fife	88	95	94	87	90	89	93	89	87	88	90	90	100	100	100	78	87	83	95	98	96	77	91	93	83	84	86	83	95	94
Glasgow City	95	89	91	99	93	94	95	88	90	88	84	87	100	100	100	98	83	83	96	81	83	85	35	43	86	24	32	86	100	0
Highland	100	100	100	49	59	53	42	54	49	53	69	46	100	100	100	43	51	46	48	54	49	48	59	53	48	56	50	48	54	49
Inverclyde	100	100	100	98	99	99	96	98	97	98	99	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Midlothian	14	13	16	86	85	86	100	99	96	84	83	83	100	100	100	47	49	50	37	36	35	45	51	52	45	50	50	45	37	36
Moray	100	100	100	91	100	100	0	16	19	100	98	100	100	100	100	93	83	100	23	100	12	100	100	100	100	100	100	100	98	100
North Ayrshire	92	94	96	93	98	99	93	97	98	96	99	100	100	100	100	27	65	100	100	95	100	100	99	100	100	98	100	100	80	92
North Lanarkshire	8.4	10	12	0	0.8	0	0	0.8	0	100	100	100	100	100	100	85	85	85	14	16	15	100	100	100	50	52	19	50	1.5	1
Orkney Islands	73	67	70	95	90	95	85	81	86	94	97	98	100	100	100	100	100	88	80	59	62	53	100	99	38	22	27	38	96	90
Perth & Kinross	100	100	99	100	99	96	99	99	88	100	100	99	100	100	100	99	100	99	99	100	99	100	100	97	100	100	91	100	99	98
Renfrewshire	100	100	100	99	99	99	99	98	96	99	100	99	100	100	100	64	73	95	100	100	100	100	99	100	17	30	100	17	17	100
Scottish Borders	100	100	100	93	90	85	92	98	98	92	90	86	100	100	100	93	93	93	100	100	100	100	100	100	76	78	78	76	100	100
Shetland Isles	50	49	48	99	77	92	97	100	100	100	93	99	100	100	100	100	59	40	100	95	99	97	100	100	74	17	100	74	100	100
South Ayrshire	100	100	100	94	94	99	100	100	100	100	100	100	100	100	100	100	100	94	93	93	100	100	100	100	98	98	82	98	81	100
South Lanarkshire	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100	100	100	94	99	100	92	99	100	92	99	100
Stirling	100	100	100	89	88	69	88	70	96	89	100	100	100	100	100	83	100	89	0	11	10	100	100	100	0	11	0	0	100	93
West Dunbartonshire	100	100	100	97	98	98	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
West Lothian	10	13	24	56	58	86	100	100	100	88	90	100	100	100	100	100	100	100	15	20	100	100	100	100	100	100	100	100	100	100

Graph 3: Autism spectrum diagnosis as percentage of non-missing data by LA, 2011-2014



4. Possible weighting schema

In the context of an administrative dataset containing missingness, an unbiased estimate can in principle be calculated as -

$\widehat{Y} = \sum_{i=1}^{n} y_i \big/ p_{yi}$, where $\widehat{Y}$ is the estimate, y_i are the observed values for a given variable for each non-missing record i, p_{yi} is the probability of y_i being non-missing for record *i*, and *n* is the number of non-missing records.

This estimator can be re-written using weights -

$$\widehat{Y} = \sum_{i=1}^{n} w_i y_i$$

$$w_i = 1 \big/ p_{yi}$$

Therefore, if we can accurately predict the probability of a record being non-missing - p_{yi} - we can produce an unbiased estimate by taking the inverse of this probability and using it as a weight. Intuitively – we are assigning larger weights to cases that are more likely to be missing.

In practice, estimating this probability is challenging. In section 3 we suggested that missingness varies between local authorities but not obviously by age, sex, or any other attribute data available for all cases. One way to estimate p_{yi} is therefore to assume constant rates of missingness within a local authority. If $n_{y,LA,non-missing}$ is the number of non-missing cases for variable *y* in an LA, and n_{LA} is the total number of cases in an LA, we can estimate -

$$p_{yi} = \frac{n_{y,LA,non-missing}}{n_{LA}}$$

The weight is then -

$$w_{yi} = \frac{n_{LA}}{n_{y,LA,non-missing}}$$

This weight is essentially ‘scaling up’ the non-missing data within a local authority to represent the missing data within a local authority. So, for example, if there were 100 adults with a learning disability in Aberdeenshire and we have data on Autism Spectrum diagnosis for 74 of them, the weight for Aberdeenshire for autism spectrum diagnosis would be 100/74=1.35. This weight ‘scales up’ the 74 cases with a response to represent all 100 cases.

However, in some cases this would produce extremely large weights – for example, in cases local authorities with only 1% completeness for a variable, the weight would be 100; and in local authorities with no response ($n_{y,LA,non-missing}$ =0) the weight is undefined. Large weights can make year-on-year estimates volatile, and raise issues with representivity – if response is only obtained from a small number of cases, it is likely to be inappropriate to use these cases to represent all cases within a local authority.

One option is to apply a maximum to the weight -

$$w_{yi} = \max\left(\frac{n_{LA}}{n_{y,LA,non-missing}}, 2\right)$$

The statistics behind setting such a maximum are complex, and depend on a number of hard-to-measure parameters. A maximum of '2' is more-or-less arbitrary, but would be reasonably straightforward to apply, and reflects the intuition that, if less than 50% of data within an LA is available, using non-missing data to represent the missing data in full may be inappropriate.

5. Options for reporting change in future

This section presents three alternatives –

- Report changes in proportions for recent years, not using any weighting method
- Report changes in proportions for recent years, using a weighting method
- Continue with the current practice of not reporting change

5.1 Report changes in proportions for recent years, not using any weighting method

The previous approach to reporting change, which in many cases focused on reporting change in totals for the entire time-series without any adjustment for missingness, should not be used. This is because reporting totals by counting non-missing respondents will systematically under-estimate true counts and may lead to spurious changes, as illustrated in section 2.

If the data were missing completely at random, it would be reasonable to report time-series for proportions and missing data could simply be excluded from these proportions – as in table 3 in section 2. Since, in this scenario, data are missing at random, we can use non-missing data to implicitly represent missing data and achieve an unbiased estimate.

However, in section 3, we showed that missingness does vary by local authorities, and learning disability outcomes also vary by local authorities. This may lead to bias in estimates of change if an estimation method which accounts for missingness is not used. For example, if individuals in a particular local authority are more likely to have an autism spectrum diagnosis, and if the missingness rate for that local authority reduces between two consecutive years, overall estimates of autism diagnosis will increase.

Because of this, if an estimation method which accounts for missingness is not used, missing data should be included in the denominator for estimates – as in table 2 in section 2. This makes interpretation of change more challenging, but is more transparent, and does not implicitly use non-missing data to represent missing data.

Bias in estimates of change will be particularly pronounced when missingness levels change substantially year-on-year. The amount of missingness in the data has dropped substantially between 2008 and 2011, as illustrated in Graph 2, but has approximately stabilised since then. Missingness for particular variables within a local authority does vary, but does appear fairly constant in recent years for most local authorities, as illustrated in graph 3.

This suggests that while reporting long-term change without adjustment is inappropriate, reporting changes in proportions for recent years without using any weighting method should produce a reasonably accurate estimate of change as long as missingness levels within most local authorities remain fairly stable. Such an approach should only be taken for estimates at the overall level and not at a regional level, as regional estimates will be particularly sensitive to changes in missingness in individual local authorities.

5.2 Report changes in proportions for recent years, using weighting scheme outlined in section 4

Section 4 presented a possible weighting schema which accounts for variation in missingness between local authorities.

An implicit assumption in this method is that missingness is at random within local authorities. If missingness is not at random within local authorities, estimates will still be biased. This is a particular concern for longer-term time series due to the drop in missingness between 2008 and 2011, and it may be preferable to focus on short-term changes even if using this weighting method.

An additional consideration is that while the proposed weighting scheme is reasonably straightforward, it would require a large number of weights to be calculated – one for each variable (due to differing rates of missingness for different variables). This is a general issue with using weights to estimate for 'item' non-response. This raises the risk of errors and would presumably reduce the amount of time available for quality-assurance and other activities. The method would also need to be applied historically, which may be challenging. It is not obvious that the benefits to using this scheme outweigh the costs.

In principle, this weighting scheme would allow more robust estimation of totals, since it removes at least some of the problem of the systematic under-representation due to missingness. However, part of this issue will still remain due to the maximum of 2 put on the weight, and we do not recommend using this method to estimate totals.

5.3 Continue with the current practice of not reporting change

The disadvantages of not reporting any change are obvious, but if statistics for change over time are not fit for purpose this is the best option.

6. Recommendations

Long-term change and estimates of change in totals should not be reported.

There is a risk in reporting short-term change in proportions without any adjustment for missingness, since changes in missingness patterns may drive changes in estimates. However, between two periods where missingness appears to have changed fairly minimally, this risk may be worthwhile for statistics where there is clear demand for estimates of change.

When reporting changes in proportions under missingness, there is a question about whether to include missing data in the denominator of the proportion, as described in section 5.1. This depends on a balance between usefulness and transparency, and whether adequate caveats can be included in the commentary. It may be preferable to include missing data in the denominator to ensure users have a full and transparent picture of how much data is missing, and accept that this reduces the utility of estimates of change.

We have presented a weighting scheme which would, in principle, remove some of the bias in estimates of change due to missingness. However, this method requires the calculation and application of a large number of weights, which would be a resource intensive and raise the risk of errors. It may be worthwhile applying the weighting scheme on an experimental basis in order to evaluate whether this estimator can reasonably be applied and the difference it makes to estimates.

Finally, it is important to emphasise that if missingess is correlated with learning disability outcomes – for example, if autism data are populated only where a positive autism spectrum diagnosis has been made - estimates of level and change will be biased under any estimation method. It is crucially important that SCLD continue to work with local authorities to understand why data are missing and ensure as far as possible that data is not disproportionately missing for some categories of outcome variables.