Handling missing data in numeric analyses

Stephen Gorard
Durham University Evidence Centre for Education
s.a.c.gorard@durham.ac.uk

Abstract

Social science datasets usually have missing cases, perhaps through non-response, and cases missing values, perhaps through dropout. All such missing data has the potential to bias any future research findings based on these datasets. This is well known, and researchers need to work to minimise missing data. However, many research reports ignore the issue of missing data, or only consider some aspects of it, or do not report how it is handled. This new paper rehearses the reasons why data might be missing, including a brief consideration of what it means for data to be missing randomly, and the damage it may cause to research. The paper then briefly considers eight different approaches to handling missing data so as to minimise that damage, their underlying assumptions and the likely costs and benefits. These approaches include complete case analysis, complete variable analysis, single imputation, multiple imputation, maximum likelihood estimation, default replacement values, weighting, and sensitivity analyses. Using only complete cases should be avoided wherever possible. The paper suggests that the more complex, modelling approaches to replacing missing data are based on questionable methodological and philosophical assumptions (e.g. that events can be caused neither randomly nor non-randomly, but somehow incompletely randomly). And they may anyway not have clear advantages over simpler approaches like default replacements. It makes sense to report all possible forms of missing data, report everything that is known about the characteristics of cases missing values, conduct simple sensitivity analyses of the potential impact of missing data on the substantive results, and retain knowledge of missingness when using any form of replacement value.

Keywords

Missing data, attrition, imputation, randomness, bias

Introduction

Missing cases

In almost all large-scale social science research, there will be valuable data missing from the dataset to be analysed (Berchtold 2019). This could be because of refusal, or non-response to an invitation to participate, leading to entire cases missing from the dataset at the outset. It could be due to dropout or attrition, where cases that had agreed to participate changed their mind, or were prevented by circumstances from providing data. Such attrition may be due to mortality, illness, busyness, or mobility of the cases. This leads to entire cases unavailable for substantive analysis, by the end of the study. Full participation in any social science study is so rare as to be unheard of (Lindner et al. 2001, Cuddeback et al. 2004). This would affect the generality of any findings to cases that had not been involved in the study.

Missing values

There will also be values missing from variables for the cases that are in the dataset. This could arise from partial non-response, refusal to answer an intrusive or complex question, indecipherable responses, loss of data, incomplete coverage of the case, movement/mobility over time, errors such as inadvertent skipping of question items, and a range of related reasons. Again, it is very unusual to find a social science dataset in which every case has complete values. This would affect the trustworthiness of the achieved results (Gorard et al. 2017).

In combination, missing cases and missing values mean two things for analysis. It is safest to assume that all datasets are incomplete. This also means that cases are never truly randomly selected, or randomly allocated to groups, because a random sample which is incomplete is no longer random (Hansen and Hurwitz 1946, Sheikh and Mattingly 1981). Any form of analysis predicated on full randomisation cannot be used where any data is missing (even if the missing values are replaced somehow – see below).

This is not a counsel of despair. The most effective and safest method for handling missing data is to try and prevent it occurring in the first place. This involves making the provision of data easy, chasing up non-responders, using memoranda of understanding with participants, and using intention to treat designs, among many other ways to inhibit loss of data. This paper focuses more on what to do after data collection, and before substantive analysis, in order to reduce the impact of whatever missing data remains.

The theoretical nature of missing data

There is a widespread idea in social science that there are three generic types or causes of missing data (Little and Rubin 2002). The first clear type is where the loss of data has occurred randomly, and so the missing data (if it could be found) would be an unbiased subset of the full dataset – both missing and non-missing. There would be no patterning of the missing data in terms of any other variables, and the reason for any data being missing would be unrelated to its value. This is largely a theoretical description. It is hard to envisage this happening for any real-life social science dataset, and is generally impossible to identify such a situation using only the data that is not missing (because if we had access to the data that is purportedly missing then it is not missing). For example, non-responders to a survey may tend to be busier, or less literate, confident or interested in the topic. This is common in social science, and it is clearly not random.

The second clear type is where the missing data has not occurred randomly, and whether it is missing or not depends on, or is linked to, the missing values themselves (as with the non-response examples above). For example, the average level of literacy of respondents might be over-estimated where less literate cases are missing from a survey dataset, because these respondents are less likely to be able to read a questionnaire. Or the estimated literacy level may be biased because of another variable in the dataset. For example, the average level of literacy may be over-estimated where homeless cases are somewhat more likely to have low literacy, and did not receive a postal questionnaire because they have no address. This can lead to both missing cases and missing values (as described above). The income of respondents may be under-estimated if wealthier respondents are more likely to omit this question for fear of tax inspection or similar. Both kinds of biases are possible wherever data is missing for any reason, and they have been found to be very common in those rare situations where it is possible to judge the patterns in the missing data (Dolton et al. 2000).

The third purported type of missing data is less clear, and is rather confusingly referred to as "missing at random", as opposed to the first type which is referred to as "missing completely at random", and the second which is "missing not at random". Here the missing data is assumed to be missing for reasons unrelated to its value, but only after controlling for some other variables. The theory is that the observed data can be used to explain or predict the missing data. Put another way, it means that any pattern in the missing data can be fully explained by looking at other variables in the dataset (Brunton-Smith et al. 2014). For example, a local authority might have records of the attainment of students at school in its area. Some students aged 16 may have moved away, and their attainment at age 16 is missing from the records. Looking at the prior attainment at age 11 of the missing cases it is clear that the missing attainment scores at age 16 will be lower than the average of the cases for whom data is available. Therefore, the missing data is not truly random. However, the "missing at random" idea is that if we control for attainment at age 11, which is strongly correlated with attainment at age 16, then attainment at age 11 explains the association between attainment at age 16 and the likelihood of missing data on

attainment at age 16. Once controlled for, the data is said to be "missing at random" (but is still not actually missing randomly). This example clearly only tries to address missing values, and is silent on missing cases.

However, a problem in a large dataset where some data is missing from one variable is that there is also likely to be missing data for other variables, making it unclear which missing values should be addressed first. Some cases might be missing figures on their attainment at age 11, and others might be missing figures on their attainment at age 16. If those missing data from age 11 are lower attaining, on average, then using age 11 figures to control for missing age 16 attainment figures might create more bias than leaving well alone. We could try and control for the missing age 11 figures as well, but this would then require a third strongly correlated variable, which might also have missing values, requiring a fourth variable, and so on.

It is much harder to envisage missing at random in relation to missing cases. As an example, if we assume that the missing cases in a survey are due to the cases being homeless (as above), and we have some homeless cases, then we might use the average literacy of the homeless cases we do have to estimate the literacy of all missing cases. This is a very unrealistic scenario, and it can never be tested for accuracy because we do not know the values of the missing data, and so cannot compare cases with and without that value to check whether they differ in terms of that variable (Soley-Bori 2013). It is also very likely that the homeless cases that did respond to the survey were different to those who did not in a way that would affect the literacy estimate. For example, if the urban homeless were more likely to respond and had higher literacy, on average, than rural homeless cases then using the figures we do have, to estimate the figures we do not have, could be ineffective or even create more bias. The assumptions underlying "missing at random" are unrealistic. For example, it would be hard to be confident that all missing data on literacy was due to homelessness (as opposed to illness, motivation or being on holiday). And it is unwarranted to assume that any homeless people who did respond had the average literacy rate for all homeless people.

There is considerable doubt, anyway, whether this supposed middle ground of missing at random (but not randomly) makes sense philosophically or mathematically. People are not good at judging and understanding randomness, and cannot identify it accurately from a finite sequence of data units (Volcan 2002). Randomness itself may not even exist (Gorard 2013a).

There have been various attempts to define randomness over time, including in terms of unpredictability, non-computability, and long-term probability of proportional selection from population characteristics (von Mises 1941). Kolmogorov and Uspenskii (1987) define a set of units as occurring randomly if the set cannot be described more efficiently than by repeating the entire set (which is perhaps another form of non-computability). None of these ideas matches the concept of "missing at random" as defined by Little and Rubin (2002). "Missing at random" values are meant to be predictable, and patterned by population characteristics, and can be summarised (or modelled by regression) more efficiently than by simply listing each one. Therefore they are not random, by any definition. None of the considered definitions of randomness allow for degrees of randomness – it is something that is or is not. Nor can figures be manipulated in order to make them random, by definition. Of course, non-random cases can appear to be more or less predictable, and easier or harder to compute, but this is not what makes them random or not. For example, the sequence 1, 2, 3, 4, 5, 6, 7…. is slightly easier to see as a pattern and to apparently predict the next item in than the sequence 1, 4, 9, 16, 25, 36, 49…. But this does not affect whether either sequence is random – something which it is impossible to tell in practice. Degrees of complexity do not imply degrees of randomness.

In summary, it is safest to assume that data is seldom missing randomly, and that there will always be a pattern created by the reason that at least some of the values are missing. All missing data therefore creates a possibility of bias in any results based on the remaining cases (Behaghel et al. 2009, Peress 2010, Hughes et al. 2019). This has been repeatedly demonstrated when population or administrative data are compared to incomplete surveys of the same cases.

The impact of missing data

It is quite common for analysts simply to ignore missing data, working with the existing cases and omitting further cases where they are missing values for a variable used in a specific analysis. This approach of working only with complete cases can create several problems, especially where we assume that the missing data is not random.

Most obviously, if the cases that are missing data for any variables are simply ignored then this reduces the number of valid cases, and so limits the appeal and trustworthiness of the analysis (Sterne et al. 2009). In an analysis that involves dividing the sample into two or more groups for comparison, there may be too few cases in any table cell for a worthwhile result. In a multivariate model such as regression, ignoring cases with any missing values can reduce the number of valid cases substantially – perhaps even to zero, in a dataset with many variables. This is so, even if the number of cases missing values for each variable is small, because any case with any missing values in any of the variables will be ignored.

Secondly, unless the data that is missing is completely random (a very unlikely scenario, see above), ignoring the whole case where any values are missing will bias the data and so produce knowingly biased results (Swalin 2018). This should be avoided. In a longitudinal study, if the poorest respondents are more likely to drop out of later waves, then estimates of average household income will become seriously inflated (Siddiqui et al. 2019). In secondary data, the most disadvantaged cases often do not have valid data even about their level of disadvantage (Gorard 2012). This then distorts both policy and practice implications from any analysis using that data. In a randomised control trial with two groups, if the more proficient cases in the treatment group find the intervention dull and are more likely to drop out, and the less proficient cases in the control group are disappointed not to be selected and are more likely to drop out, then the results will be distorted. The study may suggest that the intervention was ineffective, regardless of its actual impact, because the post-intervention scores for the control group will be artificially high, and vice versa. Examples like these can occur in any real-life study.

Third, when research-based measurements are used in calculations, any initial errors in those numbers propagate through the calculation. For example, subtracting two similar size numbers, each with errors caused by missing data or anything else, can yield a small answer that is almost entirely composed of the error parts from the original numbers. This can distort the results of an evaluation such as a randomised control trial, and is a major problem in studies of school effectiveness or similar (Gorard 2010, 2013b).

There are, therefore, good reasons not simply to ignore missing data.

A further reason that is sometimes given for dealing with missing data is that the resulting bias will affect any standard errors estimated, and so distort the significance test, confidence intervals, and other results based on these estimated standard errors. This reason is not a good one, and does not make sense. The use of significance tests and related approaches are not relevant unless the cases are fully randomised, and datasets with missing cases and values cannot be random, by definition. Nor can a dataset with missing data be made random subsequently, again by definition.

Even in the unlikely circumstances of a complete dataset in which the cases are randomly sampled (because population data and convenience samples cannot be random, again by definition), significance tests are of limited value for research. Methods experts in medicine, psychology, sociology, and education, the APA, ASA and other bodies advise against their use (Fidler et al. 2004). The American Journal of Public Health, Epidemiology, Basic and Applied Psychology, and numerous medical and ecological journals have banned their publication, as have most US medical journals (Starbuck 2016). Significance testing and p-values give misleading results about the substantive nature of results, and are 'best avoided' (Lipsey et al. 2012). Ignoring them, as we should, greatly simplifies the discussion of

handling missing data, and means that the supposed advantage of restoring the standard error, claimed by some methods, can be ignored.

Methods for handling missing data

Methods for supposedly dealing with missing data, other than ignoring it, include the use of weights, complete values analysis, multiple imputation, single imputation, maximum likelihood estimation replacement with default values, and sensitivity analyses with clear reporting (in order of decreasing complexity).

Reporting and sensitivity

A good first step, whatever other approaches are used, is to acknowledge and record all forms of missing data, the reasons for it being missing (if known), and the stage of the research at which it occurred. This then needs to be reported clearly. A useful format for many studies is a CONSORT flow diagram (http://www.consort-statement.org/consort-statement/flow-diagram). This diagram includes how many cases were approached, how many did not take part and for what reasons, how many dropped out at any stage, and how many are missing each key data element and the reasons for this, and so how many cases are and are not included in each step of the substantive analyses.

A third step is to run descriptive analyses of anything that is known about the missing cases or data. For example, in an experimental design it is important to check whether the missing data in each treatment group is balanced in terms of their known characteristics or pre-intervention scores. If the cases dropping out of one group have higher pre-test scores or are more likely to be disadvantaged or whatever, this then needs to be taken into account when providing cautions about the findings of any differences between the groups. Dropout is a bit like having missing cases (cases that do not respond not follow requests for data), and also a bit like having missing values (you have some initial data on the case, but are missing the follow up data).

Most importantly, using descriptive analyses, the researcher can judge and report the possible impact of any missing data on the substantive results. The scale of the issue might be illustrated by comparing the substantive results for all cases (using the chosen method of replacing missing values) with those for the complete cases. However, this only involves cases that have some data, and ignores the bias created by cases that are missing entirely (through non-response, for example). Where missing data is not random, the potential impact can really only be assessed through sensitivity analyses, such as estimating the most and least favourable findings by imagining favourable and unfavourable replacement values for different groups in the study (Sterne et al. 2009).

Sensitivity analyses are used in fields including economics, agriculture clinical trials, and natural sciences to support decision making, such as whether a finding is robust and so worth taking notice of substantively (Pannell 1997, Thabane et al. 2013). This simply involves looking at the impact on the substantive research findings of varying assumptions about any errors and related factors. One approach is to assess the proportion of cases that would have to be replaced with counterfactual data to invalidate the inference being made (Frank et al. 2013). When a comparison is made between the scores of two (or more) groups, it is possible to compute the number of counterfactual cases that would be needed for any difference between the group scores to disappear (Gorard 2019). For example, the counterfactual score for each group could be the mean of the other group modified to be one standard deviation away from the direction of the difference between the groups. This number of cases needed to disturb the difference (NNTD) would most easily be estimated as the overall effect size multiplied by the number of cases. If this NNTD is clearly more than the number of cases missing altogether, or the number of cases missing relevant data, then the effect size cannot have been created solely by bias due to missing data. This is a very strict and therefore robust test of the stability of any such finding, and its sensitivity to missing data.

Weighting responses

When it is clear that an achieved sample differs from the population in some crucial respect, the achieved results can be weighted to try to produce a better estimate. Weights are adjustments made to correct for the perceived bias in achieved results by using post-stratification corrections (Lehtonen and Pahkinen 1995). Weights are more often used to deal with completely missing cases than missing values.

For example, imagine collecting a stratified sample of 1,000 respondents, of whom 600 (60%) lived in urban areas and 400 (40%) lived in rural areas. However, the census of population for the region suggests that the proportion should be 80:20 urban:rural respondents. The sample over-represents rural respondents. One of the substantive questions in a survey was whether respondents had considered more than one political party when deciding on their vote at a recent election. The overall result was that 440 (44%) had considered another party, and therefore the modal average (most frequent) response is 'no' (56%). Separated by area of residence, 60% of the urban inhabitants but only 20% of the rural ones had considered voting for another party. If the achieved sample had 800 urban and 200 rural cases, and both groups had answered the voting question in the proportions achieved then 60% of the hypothetical 800 urban cases (480) and 20% of the 200 rural ones (40) would answer 'yes'. On this calculation, since area of residence makes such a difference and the sample over-represents the views of rural residents, the best estimate of the population figure considering another party might be 520 per 1000. Therefore, the modal average response is actually 'yes', even though the achieved sample appeared to suggest 'no'. This is the power and appeal of weighting.

However, weighting quickly becomes complex if several variables are involved. The underlying assumption is that missing cases would have responded in the same way as the achieved sample, but this is impossible to verify. Weights can really only be used *post hoc* to correct for variables for which all true population values are known, making weighting pointless, and weighting a sample in this way clearly cannot correct the values of other variables for which the true population value is not known (Peress 2010). It is crucial to recall that missing data of any kind must be assumed to be biased, as has been demonstrated repeatedly in practice, rather than occurring randomly. Weighting cannot overcome this. In fact, attempts at such replacement often make the bias worse. It may be safer simply to report the disproportion in the responses, to help the reader see who tentative any claims would have to be in the urban:rural example.

Complete values analysis

The previous section outlined the notion of complete case analysis, wherein all cases with missing values are dropped and computations take place only with the remaining full cases. An alternative is to keep all cases and drop only those that have missing values to be used in each analysis. So, if variables A, B and C all have missing values for different cases, then a comparison of the complete cases for A and B would include different cases to a comparison for A and C, etc. Both would involve more cases than a multivariate analysis of A, B and C, and more than a bivariate analysis based on complete cases. Despite retaining more cases for some analyses than complete case analysis, this approach is more dangerous. The number of cases obviously varies for each analysis, and so the precise subset of cases changes as well. This can be very misleading, for example, if you wanted to compare the correlation coefficients for A and B, and A and C.

Replacing with default values

A simpler approach than imputation through modelling is to replace missing values with a default that allows all cases to be included in the substantive analysis, without having an undue influence. A common default value for real number variables is the overall mean score, so that the default does not change the mean, and the cases with the default have no extreme leverage on the substantive findings. If this is done it is important also to create a new variable representing whether the original value was

missing or not. This allows analysis of the known characteristics of missing cases, and allows missingness for that variable to be used as an explanation in any substantive analysis.

A simple alternative for categorical variables is to create a further category of "missing", and use this for analysis of the known characteristics of missing cases, and as an explanation in any substantive analysis. Another approach is to recode the missing values as one of the existing categories. For example, a binary variable recording whether a respondent had attended university or not, might be recoded as whether a respondent reported attending university or not by converting missing values to not reporting attendance (the coding becomes yes or not yes, rather than yes or no). As with the use of means for real numbers, and for the same reasons, a new variable should then be created recording whether the original value was missing or not.

These approaches will tend to reduce the apparent variability of the sample, because more cases will now have the same value/category. This could be adjusted by adding/subtracting a small random component to each replacement, and the dangers of doing so are discussed below. It is better to insist that, in a large dataset, only a small fraction of values be replaced for any variable using this method. If a variable has a very high proportion of missing values then it is better to treat the whole variable as non-viable. Then, using the record of how much missing data there was originally acts as a caution to both researcher and reader against over-interpreting small "effect" sizes. This is also where a simple sensitivity analysis (above) is valuable.

These basic steps mean that all cases can be used, while allowing further analyses of missingness, and including missing as a predictive value in the substantive analysis. Used with otherwise high quality datasets, such techniques can lead to somewhat more powerful substantive models than complete case analyses (e.g. Gorard and Siddiqui 2019).

Multiple imputation

Multiple imputation is the most complex widespread approach to handling missing data. All of the variables in the dataset are used to create a model (an imputation model) that predicts or imputes the missing values for one variable. The plausible replacement values for the cases that need replacements are based on the patterns in the variables that do not have missing values. In summary, the replacement for any missing value is the value that similar cases have, who are not missing a value for this variable, plus a random element added artificially so that not all replacement values are the same. This makes the dataset complete in terms of this variable, and so the substantive research analysis can be conducted with a full dataset. This procedure is repeated many times (hence multiple) creating many full datasets by estimating different plausible replacement values each time, because of the random element in the model approach that adds variability to the predicted missing values that represent the proposed distribution of the missing data, given the data that is not missing (Brunton-Smith et al. 2014). Each complete dataset is then used to run the substantive research analysis, and the multiple results are combined to give an overall result and an estimate of variation. In reality the process is more complex than this, because it is rare for a large dataset to have only one variable with missing values. Therefore plausible values are needed for more than one variable.

Other than complexity, multiple imputation has several disadvantages for the analyst. It is not clear that even if the missing values are strongly patterned in terms of other variables their replacement values should be the same as or very similar to the values for cases not missing data (see above).

Multiple imputation is completely unable to handle missing cases (it only deals with missing values for existing cases). So, even if it worked and was justified, a major source of bias remains. The same is true of other replacement approaches, but with multiple imputation in particular this fact is often ignored as though this complicated procedure was some kind of panacea. A further limitation of multiple imputation is that its advocates say it should not be used where missing data is not random (the majority of real-life situations) or where the data is missing completely randomly. It is not clear that there can

ever be any middle ground – where data is missing just a bit randomly – and so it is not clear that multiple imputation should *ever* be used.

The identification of this missing at random status (the middle ground) cannot be determined by examining the data that does exist, and is instead merely assumed by the analyst (Crameri et al. 2015). "Missing at random" is an assumption that is used to justify the analysis, not a property of the data (Sterne et al. 2009). If the missing data is not actually at random (the middle ground), then using multiple imputation may well lead to greater bias (Hughes et al. 2019). Sterne et al. (2009) use the example of a study of the predictors of depression. If cases with depression are more likely to drop out, or miss an appointment, because they are depressed, then the necessary missing at random assumption is not viable. All of the examples above, such as longitudinal dropout patterned by household income, would be unsuitable for multiple imputation for the same reason.

It is also noteworthy that under many common conditions multiple imputation gives the same substantive results as complete case analysis anyway (such as when the missing values are in the outcome variable, or when conducting logistic regression analysis). In these circumstances, all of the complexity for researchers and readers is for no purpose.

The main argument for assuming missing at random and so being able to use multiple imputation is that any other approach can cause the associated standard errors to be smaller, because these approaches do not take uncertainty into account (Sterne et al. 2009). This could cause significance tests to produce spurious significant results. A contradictory claim is that if we do not use multiple imputation, then there is an increased chance of a Type 2 error (van Buuren 2018). But this does not matter at all if the substantive analysis to follow does not use significance tests, as it should not (see above). Therefore, the main argument for using multiple imputation no longer exists.

Modelling single imputation

A simpler alternative is to use just the first step of multiple imputation, and model the best single prediction of any missing values, adding a random element to each replacement so that not all are the same. Although simpler to conduct and comprehend, this approach basically has the same assumptions and drawbacks as multiple imputation, and does not automatically yield safer results than simpler approaches (Soley-Bori 2013, Swalin 2018).

Maximum likelihood estimation

Some commentators envisage maximum likelihood estimation as a method for replacing missing values. It involves estimating a value for a parameter which is the most likely given the available data for each case. As with multiple imputation, maximum likelihood is predicated on the missing data being "missing at random" (Dong and Peng 2013). But, as shown above, missing at random as opposed to randomly is not a viable description, mathematically or philosophically. Maximum likelihood is also really only applicable for specific types of statistical modelling, and does not handle generic missing data, as such.

Conclusion

In social science, it is still the case that too many research reports ignore the issue of missing data, not reporting it at all, or not reporting clearly how missing data is handled (Berchtold 2019). Even where research reports do cover missing data, it is not clear that missing data has always been handled appropriately. In most real-life datasets there will be missing cases and missing data, and for most of these the data will not be missing randomly. This means that complete case analyses, excluding cases with any data missing, or complete values analysis, excluding missing values, will produce somewhat biased substantive results. With non-random missing data, default replacement values can be used with

care, as long as any knowledge of which missing cases are missing data is retained, and used for further investigation.

Analysts attempting to deal with missing data do so for a number of reasons. They may want to retain as many cases as possible, and to create substantive results that are as unbiased as possible (or least not be misled by bias). However, as this paper shows, these aims will not generally be achieved using complex approaches, chiefly because the underlying assumptions are so rarely met in real life. There is considerable doubt that bias in substantive results caused by missing data can ever be corrected by technical means (Cuddeback et al. 2004). It makes more sense to report all possible forms of missing data, report everything that is known about the characteristics of missing cases and cases missing values, conduct simple sensitivity analyses of the potential impact of these missing data on the substantive results, and retain knowledge of missingness when using default replacement values so that this can be used in the substantive analyses.

References

Behaghel, L., Crepon, B., Gurgand, M. and Le Barbanchon, T. (2009) *Sample attrition bias in randomized surveys: a tale of two surveys*, IZA Discussion Paper 4162, http://ftp.iza.org/dp4162.pdf, accessed 060714

Berchtold, A. (2019) Treatment and reporting of item-level missing data in social science research, *International Journal of Social Research Methodology*, 10.1080/13645579.2018.1563978

Brunton-Smith, I., Carpenter, J,, Kenward, M. and Tarling, R. (2014) Multiple imputation for handling missing data in social research, *Social Research Update*, 65, Autumn 2014

Crameri A., von Wyl, A., Koemeda M., Schulthess P. and Tschuschke V. (2015) Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy, *Frontiers in Psychology*, 6, 1042, 10.3389/fpsyg.2015.01042

Cuddeback, G.. Wilson, E., Orme, J. and Combs-Orme, T. (2004) Detecting and statistically correcting sample selection bias, *Journal of Social Service Research*, 30, 3, 19-30

Dolton, Lindeboom, M. and Van den Berg, G. (2000) *Survey Attrition: A taxonomy and the search for valid instruments to correct for biases*, http://www.fcsm.gov/99papers/berlin.html

Dong, Y. and Peng, C. (2013) *Principled missing data methods for researchers*, Springer Open, 10.1186/2193-1801-2-222

Fidler, F., Thomson, N., Cumming, G., Finch, S. and Leeman, J. (2004) Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine, *Psychological Science***,** 15, 2, **119-126**

Frank, K., Maroulis, S., Doung, M. and Kelcey, B. (2013) What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences, *Educational Evaluation and Policy Analysis*, 35, 4, 437-460

Gorard, S. (2010) Serious doubts about school effectiveness, *British Educational Research Journal*, 36, 5, 735-766

Gorard, S. (2012) Who is eligible for free school meals?: Characterising FSM as a measure of disadvantage in England, *British Educational Research Journal*, 38, 6, 1003-1017

Gorard, S. (2013a) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2013b) The propagation of errors in experimental data analysis: a comparison of pre- and post-test designs, *International Journal of Research and Method in Education*, 36, 4, 372-385

Gorard, S. (2019) Do we really need Confidence Intervals in the new statistics?, *International Journal of Social Research Methodology*, 22, 3, 281-291

Gorard, S. and Siddiqui, N. (2019) How trajectories of disadvantage help explain school attainment, *SAGE Open*, https://journals.sagepub.com/doi/10.1177/2158244018825171

Gorard, S., See, BH and Siddiqui, N. (2017) *The trials of evidence-based education*, London: Routledge

Hansen, M. and Hurwitz, W. (1946) The problem of non-response in sample surveys, *Journal of the American Statistical Association, 41*, 517–529

Hughes, R., heron, J., Sterne, J. and Tilling, K. (2019) Accounting for missing data in statistical analyses: multiple imputation is not always the answer, *International Journal of Epidemiology*, https://doi.org/10.1093/ije/dyz032

Kolmogorov, A. and Uspenskii, V. (1987) Algorithms and randomness, *Theory of Probability and its Applications*, 32, 3, 389-412

Lehtonen, R. and Pahkinen, E. (1995) *Practical methods for design and analysis of complex surveys*, Chichester: John Wiley and Sons

Lindner, J., Murphy, T. and Briers, G. (2001) Handling non-response in social science research, *Journal of Agricultural Education*, 42, 3, 43-53

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences

Little, R., and Rubin, D. (2002) *Statistical Analysis with Missing Data, 2nd Edition*, New York, NY: Wiley

Pannell, D. (1997) Sensitivity analysis of normative economic models: Theoretical framework and practical strategies, *Agricultural Economics*, 16, 139-152

Peress, M. (2010) *Correcting for Survey Nonresponse Using Variable Response Propensity*, Journal of the American Statistical Association, http://www.rochester.edu/College/faculty/mperess/Nonresponse.pdf

Sheikh, K. and Mattingly, S. (1981) Investigating nonresponse bias in mail surveys, *Journal of Epidemiology and Community Health, 35*, 293–296

Siddiqui, N., Boliver, V. and Gorard, S. (2019) Assessing the reliability of longitudinal social surveys of access to higher education: the case of the *Next Steps* survey in England, *Social Inclusion*, 7, 1, DOI: 10.17645/si.vXiX.1631

Soley-Bori, M. (2013) *Dealing with missing data: Key assumptions and methods for applied analysis*, Technical Report No. 4, Boston University School of Public Health, https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

Starbuck, W. (2016) 60th Anniversary Essay: How journals could improve research practices in social science, *Administrative Science Quarterly*, 61, 2, 165–183

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A. and Carpenter, J. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *BMJ*, 338, b2393, 10.1136/bmj.b2393

Swalin, A. (2018) *How to handle missing data*, https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Thabane , L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M. et al. (2013) A tutorial on sensitivity analyses in clinical trials, *BMC Medical Research Methodology*, 13, 92, http://www.biomedcentral.com/1471-2288/13/92

Volcan, S. (2002) *What is a random sequence?*, The Mathematical Association of America, 109, https://www.maa.org/sites/default/files/pdf/upload_library/22/Ford/Volchan46-63.pdf

van Buuren, S. (2018) *Flexible imputation of missing data*, Chapman and Hall

von Mises, R. and Doob, J. (1941) Discussions of papers in probability theory, *Annals of Mathematical Statistics*, 12, 2, 1941, 215-217