Method Article

An absolute deviation approach to assessing correlation

Abstract

This paper describes two possible alternatives to the more traditional Pearson's R correlation coefficient, both based on using the mean absolute deviation, rather than the standard deviation, as a measure of dispersion. Pearson's R is well-established and has many advantages. However, these newer variants also have several advantages, including greater simplicity and ease of computation, and perhaps greater tolerance of underlying assumptions (such as the need for linearity). The first alternative approach simply divides the co-variance by the absolute mean deviation(s) instead of the standard deviation as in Pearson's R. The second alternative uses the sum of each pair of deviations in x and y instead of the covariance, and again uses the absolute mean deviation(s) as the denominator. All three are compared to one another using 30,000 simulations based 100 pairs of random numbers. The substantive findings are the same for each approach, and the 'coefficients' correlate with each other (using R) at +0.99 to 1.00. The three approaches also give the same substantive findings when trialled with real-life secondary datasets. The introduction of simpler kinds of regression forms part of an attempt to simplify the use of numeric analysis, to make it more 'everyday', for the benefit of both analysts and consumers of evidence.

Introduction

In the UK, there is a long-standing concern about the poor quality and utility of social science and public policy research, and this poor quality has been largely attributed to deficiencies in methods (McIntyre and McIntyre 2000, Schuller 2007, Goldacre 2013). Similar concerns have arisen in other countries (National Science Foundation 2002). Recent capacity-building efforts have focused on a purported lack of work using numbers, and the solution proposed is a national Quantitative Methods Initiative for social science (Economic and Social Research Council 2014). That initiative is funded by a range of partners including the Higher Education Funding Council for England, the British Academy, the Nuffield Foundation, and the Economic and Social Research Council. The focus is on the development of new methods courses, teaching resources, new training opportunities, work placements, scholarships, and the embedding of 'quantitative' evidence in substantive courses for all upcoming social scientists. The model underlying all of this activity and expenditure is that 'quantitative' work in social science is currently good, especially in economics and psychology, but that it is not always well taught, and there is just not enough of it outside those two areas (MacInnes 2009, p.2).

Some of this diagnosis may be correct, although it is possible that the problems of social science are as much to with lack of design, genuine curiosity and research integrity as to do with there not being enough studies that involve numbers *per se*. Many of the proposed solutions also may be necessary but they will probably not be sufficient. What is being taught as 'quantitative' methods is currently too often wrong (the meaning of significance tests is widely mis-described, for example – see Gorard 2010), and even more often needlessly complex (the use of multi-level modelling with population data, for example). It is these

confusing approaches more than anything that are driving researchers away from the relatively simple use of measurements and frequencies in their work (Cooper and Shore 2008, Hampel 1997, Murtonen and Lehtinen 2003, Watts 1991). An alternative approach is to simplify what it means to use numbers in research, to stress the solid links between the logic of analysis for any kind of data, and then try to improve uptake through the pedagogy and persuasion of something like the Quantitative Methods Initiative. This paper is one step in that approach.

The paper introduces two possible alternatives to the more traditional Pearson's R correlation coefficient, both based on using the mean absolute deviation, rather than the standard deviation, as a measure of dispersion. These variants have several advantages, including greater simplicity and ease of computation, and greater tolerance of underlying assumptions (such as the need for linearity). The paper forms part of an attempt to simplify the use of numeric analysis, to make it more 'everyday', for the benefit of both analysts and the consumers of evidence.The approach includes replacing the standard deviation (and its needless squaring and square rooting) with the absolute deviation where possible, introducing a robust absolute deviation effect size (Gorard 2013, 2014a), and absolute deviation regression models, and of course the removal of significance testing and all of its components from everyday analysis (Gorard 2014b).

Correlation is a measure of the strength of a relationship between two variables, where each value in one variable has a corresponding or paired value in the other variable. An example might be a list of test scores in maths and reading for the same children. Correlation is the key idea underlying more complex modelling including forms of regression and data reduction. It is usually based on a scatterplot showing a linear or near-linear relationship between the two variables, and the correlation 'coefficient' is an estimate of how well the pairs of values in both variables 'fit' a straight line (Gorard 2003). In this type of correlation, the values must be real numbers (and there are other techniques for categorical data). The traditionalcorrelation coefficient is Pearson's R.

Pearson's R correlation coefficient

Correlation generally starts with covariance. 'Covariance' is an estimate of the scale of inter-relatedness of two variables. For a set of cases involving two variables x and y, we can find the deviation of each value from the mean of that variable, multiply the results for each pair of values, and sum these products for all values, before dividing by the number of values. This covariance, or average co-deviation, is large where each value in a pair differs from its mean in the same direction consistently.

$$[\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})]/n \qquad \text{(Covariance)}$$

The covariance is standardised and adjusted for the scale of measurement by dividing it by the product of the two standard deviations for the two variables. This produces the correlation coefficient known as R.

$$\frac{\left[\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})\right]/n}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2/n\right]}.\sqrt{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2/n\right]}} \qquad \text{(Pearson's R)}$$

The numerator is the covariance of the two sets of numbers x (from 1 to n) and y (from 1 to n).This covariance is the sum of: each numbers' deviation from the mean of one set multiplied by the equivalent numbers' deviation from the mean of the other set, all divided by the number of numbers in both sets. The denominator is the product of the standard deviations of the x and y sets of numbers individually. Although usually invisible to analysts using software like SPSS this is quite a complex formula. It is not an easy formula to explain to general under-graduates in social science, for example. The squaring and square-rooting involved in calculating the standard deviation is not intuitive, and has the effect of over-emphasising larger deviations from the mean. There are simpler alternatives.

Using the mean absolute deviation

Originally the standard deviation (SD) was devised as a way of eliminating the signs for each deviation. By definition, the sum of all of the deviations from the mean for any set of numbers will be zero. Squaring them before summing makes them all positive, and so the sum of all deviations would only be zero if every number in the set was equal to the mean (i.e. there were no deviations).

$$\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2/n\right]} \qquad \text{(Standard deviation)}$$

A simpler alternative to the standard deviation, as a measure of dispersion, is the mean of the absolute deviations from the mean (|MD|). Here the sum is of the absolute values (irrespective of sign) of each deviation, and because this is divided by the number of values, it is the mean absolute deviation. It gives the same substantive result as the standard deviation because the two are, of course, closely related (Gorard 2014a).

$$\sum_{i=1}^{n}|x_i - \bar{x}|\,/n \qquad \text{(Mean absolute deviation)}$$

This is the average by which each value deviates from the mean in each set of numbers (as opposed to the average deviation from the median of the numbers, which could also be used – El Amir 2012b). Such an everyday meaning is lost in the more complex calculation of the standard deviation, because of the squaring and square-rooting. And that is part of the problem when introducing a measure of dispersion like the standard deviation to new researchers. This ease of understanding is only one reason for increasingly preferring the

absolute deviation over the standard deviation. The absolute deviation is also more robust when dealing with real-life data not necessarily following a normal distribution (Gorard 2005). In fact, its use requires no assumptions about the format of the data other than that it is based on real numbers.

There is anyway a widespread but apparently disregarded ambiguity in the use and reporting of standard deviations. The formula for calculating a standard deviation makes it clear that the sum of squares is divided by n, and then the square root of the result is taken. Any square root for a real positive number must always consist of two values (one positive and one negative). In practice, however, the positive square root is often the only one quoted and considered by analysts. Statistical software such as SPSS and office software such as Excel, for example,only ever displays the positive or unsigned result as the standard deviation. So it is only the positive square root that is generally propagated into future calculations (like calculating a standard deviation). The formula should really be adjusted to portray that it is the absolute value of the square root that is being used. But then the main reason why the standard deviation was devised was to evade the use of absolute values, which are inconvenient in many ways for algebraic manipulation. In practice, the use of what are clearly absolute values in the calculation of standard deviations means that any advantage they may have had over absolute deviations now disappears completely.

Using the absolute value of the standard deviation, as happens in practice, can alsocause ‘errors’, since in some calculations half of the possible ensuing outcomes will be ignored. A widespread example is an effect size such as Cohen’s d, where a difference between means is standardised through division by their standard deviation(s). The result should therefore really be both positive and negative (because the standard deviation in the denominator should be both positive and negative). This would mean that it is not possible to tell from the effect size alone in which direction it operates. For example, it would not portray whether the experimental or control group had performed better in a simple randomised controlled trial. Of course, it is possible to tell this by other means (such as inspecting the data), but the summary effect size itself would be neutral on the matter, even using the convention of putting an intervention group before the control group score. Therefore, a simpler form of analysis based solely on absolute deviations could be presented instead. This is the idea examined further in this paper.

An absolute deviation correlation coefficient

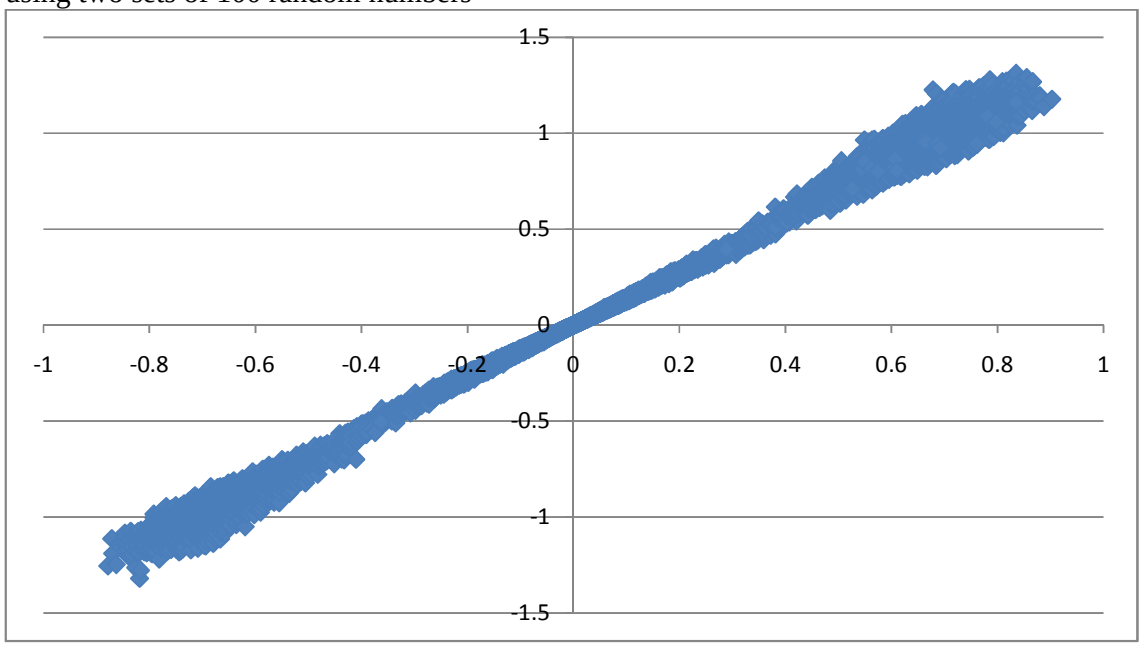
One simpler alternative for a correlation coefficient would be to replace the standard deviations in the denominator of R with the mean absolute deviation of each set of numbers instead. This would be:

$$\frac{\left[\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})\right]/n}{\sum_{i=1}^{n}|x_i - \bar{x}|/n.\sum_{i=1}^{n}|y_i - \bar{y}|/n} \qquad \text{(new correlation coefficient RA1)}$$

This approach does not involve squaring the deviations of each value from the mean, summing the results and then finding the square root. The resulting formula is

thereforesimpler than for Pearson's R. It is the sum of the products of each pair of deviations divided by the product of the sums of each pair of deviations. As a measure of dispersion, the mean absolute deviation is robust and efficient. Unlike the standard deviation, it does not artificially and unnecessarily inflate the larger deviations by squaring them, and then incompletely square-rooting the results. In almost all other practical circumstances it does not matter whether the standard or absolute deviation is used. For example, creating 30,000 pairs of samples, each of 100 random numbers, allows the correlation between each pair to be assessed. Doing this with both Pearsons's R and the absolute deviation correlation (here termed 'RA1') yields two sets of 30,000 examples (Figure 1). This shows that there is a clear and straight-line equivalence between the two measures of correlation. Using R, their inter-'correlation' is reported as +1.0.

Figure 1 – Scatterplot of values for R (x axis) and RA1 (y axis), each based on 30,000 trials using two sets of 100 random numbers



A standard deviation is generally larger than an absolute deviation for the same set of figures (because it inflates the figures slightly via the squaring and square-rooting process). This means, in turn, that the absolute deviation correlation coefficient (created by dividing the covariance by the absolute deviations) is generally larger in absolute terms than R for the same two sets of figures. RA1 is usually greater than R by a factor of around 1.3, because the absolute deviation is usually smaller than the standard deviation by a factor of about 0.77. Some commentators have stated that there is a simple conversion from SD to M|D|, through multiplying by $\sqrt{(2/\pi)}$ or 0.77. In fact this is not so. Each SD can have more than one M|D|, and vice versa, again caused by the process of squaring the deviations, summing and then square-rooting. And this is why the graph in Figure 1 appears to divergesomewhatfrom a straight line towards the extremes. Nevertheless, either coefficient will work as well as the other in practice, and the 0.77 or 1/0.77 conversion will usually work well enough to see what the other coefficient would have been. Before comparing the strengths and weaknesses of each approach more fully, it is interesting to see if the coefficient can be simplified even further.

An additive coefficient

In some respects it seems obvious to create a correlation coefficient that matches Pearson's R but with the standard deviations replaced by absolute deviations (as above). It might also be possible to create a kind of correlation coefficient in which the deviations for each pair of numbers are summed, and are divided by the sum of the sums of the deviation for each number separately. Where both numbers in a pair deviate in the same direction they will add together, and when in the opposite direction they will tend to cancel out. The sum of these combined deviations for both pairs together can be assessed as a proportion of the sum of the absolute deviations for each set of numbers alone.

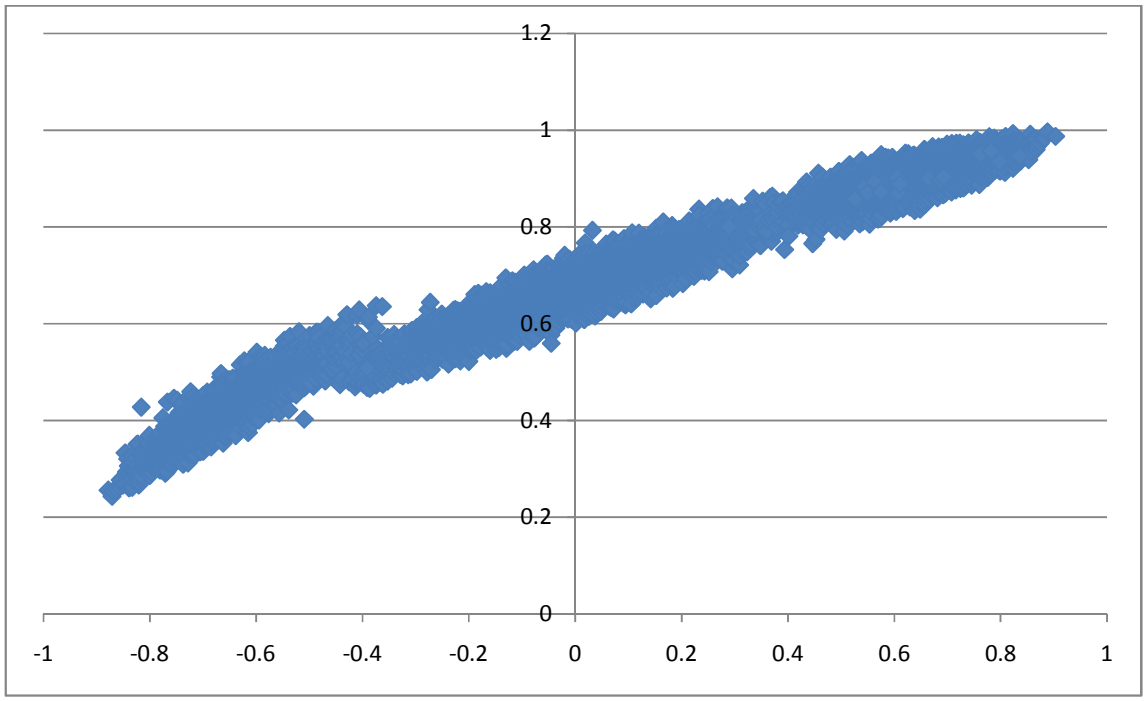
$$\frac{\sum_{i=1}^{n}|(x_i - \bar{x})+(y_i - \bar{y})|/n}{\sum_{i=1}^{n}|x_i - \bar{x}|/n+\sum_{i=1}^{n}|y_i - \bar{y}|/n} \qquad \text{(new correlation coefficient RA2)}$$

And,unlike the more complex Pearson's R, the value of n can be cancelled in numerator and denominator. This will yield:

$$\frac{\sum_{i=1}^{n}|(x_i - \bar{x})+(y_i - \bar{y})|}{\sum_{i=1}^{n}|x_i - \bar{x}|+\sum_{i=1}^{n}|y_i - \bar{y}|} \qquad \text{(new correlation coefficient RA2)}$$

This is even simpler again, and therefore should be even easier for new researchers to learn and use. As with RA1 the results from using RA2 again correlate very highly with those for Pearson's R. Here the results are illustrated for 30,000 pairs of samples, each of 100 random numbers. Running this simulation with both Pearsons's R and the absolute deviation correlation coefficient (RA2) yields two sets of 30,000 examples (Figure 2). This shows that there is a clear and straight-line equivalence between the two measures of correlation. Using R, their 'correlation' is reported as over +0.99. The use of absolute values in both top and bottom of the formula means that RA2 is always positive (or zero). Zero for RA2 is equivalent to -1 for R. Zero for R is equivalent to 0.667 for RA2 (where the regression line crosses the y axis in Figure 2). When R is 1, RA2 is also 1. What this might mean is discussed in the final section.

Figure 2 – Scatterplot of values for R (x axis) and RA2 (y axis), each based on 30,000 trials using two sets of 100 random numbers



A real-life comparison

It is also interesting to observe the behaviour of these two new proposed correlation approaches using real datasets. One such analysis is based on the annual school-level census of secondary schools in England. The figures for the characteristics of pupils in each school can be used to assess how clustered within specific schools, or 'segregated' between schools, are those pupils with any indicator of potential disadvantage. This is done using a segregation index, and the detailed method is reported in Gorard and See (2013). This means that in each year it is possible to say how segregated the school system is in terms of poor children (those eligible for free school meals or FSM), children with a disability or special educational need (SEN), and children from a minority ethnic background (non-White). It is clear that the levels of segregation and the changes over time are different for these different indicators (Gorard et al. 2013). While segregation by poverty has risen since 1998 and is now falling again, segregation of SEN pupils has dropped over time and is now rising slightly, and segregation by ethnic origin has fallen continuously. This picture yields the correlations in Table 1.

Table 1 –Comparing three correlation 'coefficients' for the same datasets: correlating the unintended segregation of FMS, SEN and non-White pupils in secondary schools in England, 1998-2014

Correlation approach	R	RA1	RA2
Segregation FSM and SEN pupils	-0.85	-1.09	0.26
Segregation SEN and non-White pupils	+0.61	+0.79	0.94

Table 1 shows that each coefficient remains related to each other in the same way as in the simulations above. R and RA1 are positive or negative together, and R is between 0.77 and 0.78 times RA1 for both figures (see above). They are both measuring the same thing and

yielding the same substantive results. RA2 does not have actual negative values, but as Figure 2 shows any value below 0.66 can be regarded as showing a negative relationship. Both values of RA2 in Table 1 could be read from the corresponding value of R in Figure 2. Again RA2 is yielding the same substantive results as R, but perhaps is a slightly harder to interpret form.

Comparing the three approaches

What this brief paper has shown is that it is possible to create alternatives to Pearson's R that are to some extent simpler (algebraically), based on the absolute mean deviation (see also El Amir 2012b). In practice, and as long as all three coefficients are understood, which is used in practice would make little difference (and many more trials have been conducted than are reported in the Figures and Tables here). They all give the same, or very nearly the same, substantive result in the contexts presented here. In many respects also they all have similar characteristics.

Pearson's R has the advantages of being built into software, its current use in other contexts, and its familiarity for existing users, especially concerning its limits, scale and importance in any context. The latter is important given that significance testing and its derivatives are inappropriate in most real-life research contexts (Berk and Freedman 2001), and do not provide an accurate probability of the 'significance' of a correlation even when all assumptions are met (Gorard 2010). What is needed instead is judgement based on the scale of the correlation (the effect size), the number of cases, and the quality of data (Lipsey et al. 2012). Familiarity helps such judgement. R also has the advantage over RA2 of having a zero that portrays easily when there is no linear covariance. It is symmetrical,as long as it is remembered that this is based onusing the absolute values of the standard deviations instead of the two (both positive and negative) versions of the standard deviations themselves. Perhaps most importantly, it is clearly and conveniently bounded by -1 and +1.

RA1 is very similar to R as would be expected where the covariance is being divided by two closely related estimates of dispersion. Clearly, when x and y are unrelated the covariance or denominator for both coefficients is zero, and thus both coefficients are zero. Therefore RA1 shares this advantage with R. However, RA1 is generally simpler to explain to new researchers,and slightly easier to calculate. Based on the mean absolute deviation, it will be more robust in the majority of real-life situations where the data contains errors and/or is not perfectly normally-distributed (Gorard 2005).Where the dataset consists of ordinal values rather than real numbers, Spearman's rank correlation coefficient is a useful alternative having the distribution-free appeal of RA1 and the convenience of R (Siegel 1956). However, using Spearman's rho where the dataset consists of real numbers means discarding a considerable amount of information – treating real numbers as mere ranks – which also means that the results for rho correlate less highly with R than those for RA1 and RA2 do.

RA1 is less distorted than R by larger deviations due to the lack of squaring, and so does not promote the deletion of inconvenient data by unwary analysts. It also appears more tolerant of non-linearity. As a scatterplot diverges from a linear crossplot of x and y but the two sets of data retain a clear relationship, such as when one is a power of the other, R decreases rapidly. In these circumstances, RA1 does not decrease. The real-life correlation of x with x+1 is no stronger than the correlation of x with x^2, for example. R merely makes it appear to be so, because x is not clearly *linearly* related with x^2. This stable and robust characteristic of

RA1 is one reason why the factor of 1.3 (above) is only a guide for when x and y are linearly related. Using RA1 early in an investigation might help pick up indications of a more complex relationship than a simple linear one between x and y.

RA2 is different to both R and RA1 in some respects. Based on addition not multiplication, and on absolute values in both numerator and denominator, it does not have a negative and positive form. In this respect it is no different to traditional effect sizes calculated properly (see above). It might therefore be necessary to make a cursory examination of the data or look at a crossplot in order to be certain of the direction of any relationship. The major unique advantage of RA2 is its conceptual and arithmetic simplicity. Like RA1 it is more robust, less distorted by large deviations, and more tolerant of non-linearity than R.

Where the results from each approach differ somewhat, as they do at extremes in Figure 1 and throughout Figure 2, it not immediately clear which result is preferable. To the extent that the different coefficients yield the same substantive results (as in measuring a temperature in either Centigrade or Fahrenheit) then scientifically it does not matter which one is used. Where they differ, given that the distortion caused by squaring is already well-known, there is an argument that the absolute deviation coefficients might more accurately portray what is being estimated.

Clearly more work would be needed than presented in this introductory paper to make either RA1 or RA2 something that could be used safely in practice, and offered as a genuinely simpler alternative to R. Analysts would also need more practice in assessing the meaning of a new coefficient value. It is not as easy, for example, as using the mean absolute deviation to create a new effect size (Gorard 2013). But the work so far shows that in principle the mean absolute deviation can be used instead of the standard deviation in a variety of settings, probably throughout descriptive statistics (Gorard 2014a). The overall objective is to make working with numbers as simple as possible for reluctant social scientists. At present, in the UK at least, considerable resources are being spent on trying to widen the use of 'quantitative' methods. Almost all of the effort lies in persuasion and pedagogy. These might be necessary, but will not be sufficient in themselves. This paper is part of an attempt to point in a different way – correcting the logical errors and simplifying what it is deemed necessary to know, before then improving the teaching of the corrected and simpler methods.

References

Berk, R. and Freedman, D. (2001) Statistical assumptions as empirical commitments, http://www.stat.berkeley.edu/~census/berk2.pdf, accessed 030714

Cooper L. and Shore F. (2008) Students' Misconceptions in Interpreting Center and Variability of Data Represented via Histograms and Stem-and-leaf Plots, *Journal of Statistics Education*, 16, 2, 13 pages

Economic and Social Research Council (2014) Quantitative Methods Initiative, http://www.esrc.ac.uk/research/skills-training-development/qmi/index.aspx, accessed 23/5/14

El Amir, E. (2012) Mean absolute deviation about median as a tool of explanatory data analysis, *International Journal of Research in Arts and Social Sciences*, 11, 3, http://www.arpapress.com/Volumes/Vol11Issue3/IJRRAS_11_3_17.pdf

El Amir, E. (2012) On uses of mean absolute deviation: decomposition, skewness and correlation coefficients, *Metron: International Journal of Statistics*, LXX, 2-3.

Goldacre, B. (2013) Building evidence into education, DfE, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/193913/Building_evidence_into_education.pdf, accessed 23/5/14

Gorard, S. (2003) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum

Gorard, S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation, *TheBritish Journal of Educational Studies*, 53, 4, 417-430

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77

Gorard, S. (2013) The possible advantages of the mean absolute deviation 'effect' size, *Social Research Update*, 65, Winter 2013, pp.1-4, http://sru.soc.surrey.ac.uk/SRU64.pdf

Gorard, S. (2014a) Introducing the mean absolute deviation 'effect' size, *International Journal Research and Methods in Education*, (forthcoming)

Gorard, S. (2014) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10

Gorard, S. and See, BH. (2013) *Overcoming disadvantage in education*, London: Routledge

Gorard, S., Hordosy, R. and See, BH. (2013) Narrowing the determinants of segregation between schools 1996-2011, *Journal of School Choice*, 7, 2, 182-195

Hampel, F. (1997) *Is statistics too difficult?*, Research Report 81, Seminar fur Statistik, Eidgenossiche Technische Hochschule, Switzerland

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

MacInnes, J, (2009) *Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK*, Report of the Strategic Adviser for Quantitative Methods to the Economic and Social Research Council, http://www.esrc.ac.uk/_images/undergraduate_quantitative_research_methods_tcm8-2722.pdf, accessed 180714

McIntyre, D. and McIntyre, A. (2000) Capacity for research into teaching and learning, Swindon: Report to the ESRC Teaching and Learning Research Programme

Murtonen, M. and Lehtinen, E. (2003) Difficulties experienced by education and sociology students in quantitative methods courses, *Studies in Higher Education*, 28, 2, 171-185

National Science Foundation (2002) *Scientific research in education*, Washington DC: National Academy Press

Schuller, T. (2007) Capacity building in educational research: sketching an international picture, http://www.scotedreview.org.uk/pdf/198.pdf, accessed 23/5/14

Siegel, S. (1956) *Nonparametric statistics for the behavioural sciences*, Tokyo: McGraw Hill

Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291