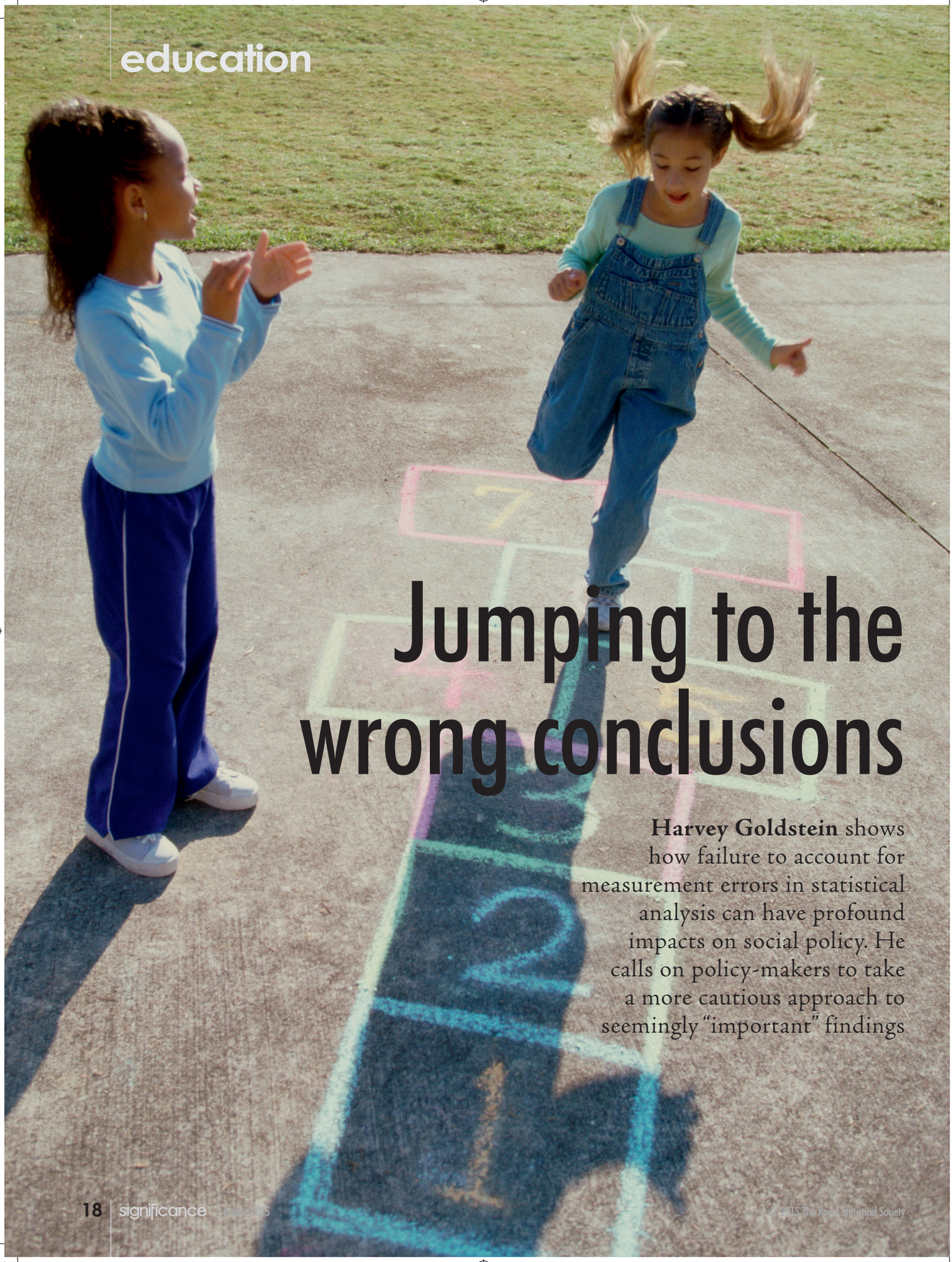

Jumping to the wrong conclusions

Harvey Goldstein shows how failure to account for measurement errors in statistical analysis can have profound impacts on social policy. He calls on policy-makers to take a more cautious approach to seemingly "important" findings

Statistical methods, by default, assume perfectly measured data are fed into them, but in practice this rarely happens, which can make our results incorrect. We call this "measurement error": instead of observing the true value X, we instead observe W, which equals X plus some measurement error U.

If errors of measurement are not properly adjusted for, they can bias the estimated values of parameters. Which is to say: we might draw the wrong conclusions from a data set, which can have important implications.

To demonstrate this, we will consider an example from the field of education policy. In 2003 Leon Feinstein, then an academic at the London School of Economics and subsequently an adviser to the UK government's Cabinet Office, published an analysis in which he claimed to show an increasing inequality in educational attainment between children from low- and high-income families.[1] His results are illustrated in Figure 1.

The data are derived from tests given to children, which sought to measure developmental status but are usually referred to as "ability" scores. In Figure 1 the circles, for example, show the mean test scores of low-scoring children who are classified as having a high socio-economic status (SES), while the triangles represent those of the low-scoring, low-SES children. Children who were below the 10th percentile of the test score distribution at the age of 22 months were classified as low scorers; those above the 90th percentile at the same age were classified as high scorers.

If you compare the circles and triangles in Figure 1, you can see that the test scores of initially low-scoring, high-SES children increase steadily over time, while those of low-scoring, low-SES children remain relatively low after 40 months.

This work by Feinstein has been quoted extensively by the media and many policy-makers to justify the preferential targeting of resources at low-SES children in the early years of life. Thus, in 2011, for example, then Deputy Prime Minister Nick Clegg announced in a House of Commons debate the launch of the UK government's social mobility strategy. Clegg claimed that: "By the age of five, bright children from poorer backgrounds [the diamonds in Figure 1] have been overtaken by less bright children from richer ones – and from this point on, the gaps tend to widen still further."

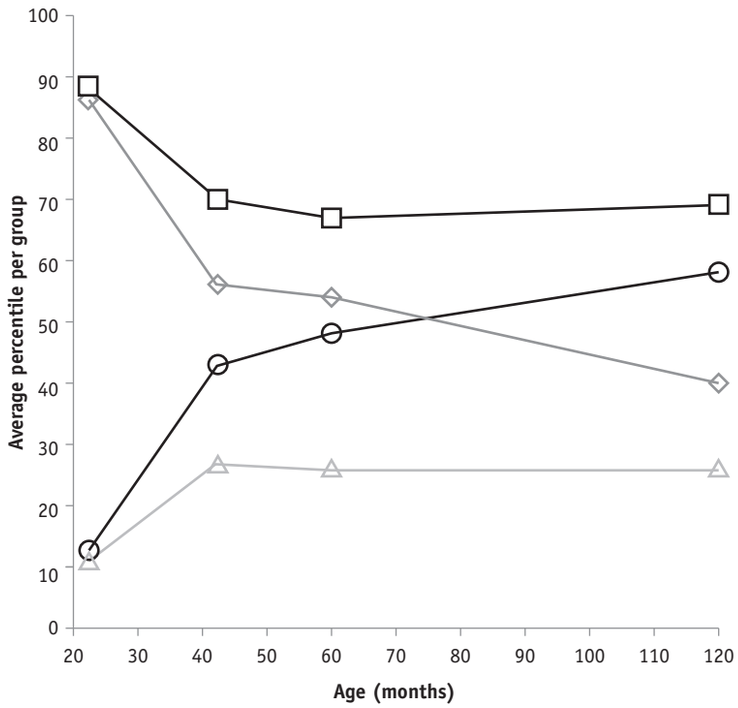

Figure 1. Development of high- and low-ability children by socio-economic group: evidence from the existing literature. ◇ high ability, low socio-economic status (SES); □ high ability, high SES; ○ low ability, high SES; △ low ability, low SES. Adapted from Feinstein,[1] based on 1970 British cohort study data

Image: Purestock/Thinkstock

Despite the widespread acceptance of these results, there are some who have been critical of Feinstein's analysis, most notably Jerrim and Vignoles.[2] The principal objection has been that the measure of ability (developmental status) at age 22 months is unreliable in the sense that a non-trivial component is effectively random noise superimposed on an underlying "true value" (see box on page 20). If one corrects for this, they claimed, a very different picture emerges.

In addition, the actual distribution of test scores within the lowest-scoring group is not the same for the two SES groups. In particular, the high-SES children have an average score that is higher than that of the low-SES group, and hence would be expected to perform better subsequently, even if there were no real relative change between the two groups.

Clarifying terms

To be fair to Feinstein, he is far from alone in ignoring the effects of measurement unreliability, but it is important both to understand the implications of this and to encourage data analysts to take it seriously. In this article I shall present some results of a reanalysis of the data used by Jerrim and Vignoles from the Millennium Cohort Study to illustrate the problem. A more detailed discussion can be found in Goldstein and French.[3]

First of all, it is important to clarify the terms being used in this debate. One term that has been used, extensively but confusingly, is that of "regression to the mean". This notion was introduced by Francis Galton for the situation where there is a less than perfect correlation between two measurements over time, as is the case with heights of fathers and their sons. For example, for tall fathers, say at the upper 95th percentile, the average adult height (measured without error) of their sons will be below the 95th percentile of the height distribution of sons – hence the term "regression to the mean". Likewise, the sons of short(er) men will, on average, be taller than their fathers.

The notion of measurement error, however, is entirely separate from that of regression to the mean, and refers instead to imperfectly measured variables. It is of course true that if one had a pair of "true" measures that were perfectly correlated (that is, with a

Measurement error

As a simple illustration of measurement error, consider the simple regression model, where the observed predictor is x_{obs} and modelling using this gives

$$y = a^* + b^* x_{obs} + e^*$$

The model we would really like to fit is the one that uses the "true" value of the predictor

$$y = a + b x_{true} + e$$

where the two are connected by a simple model

$$x_{obs} = x_{true} + \delta$$

where δ is assumed to be a random measurement error.

By "true value" is meant, roughly, the long-run mean of the predictor where it is (notionally and independently) measured in different contexts.

If we know the reliability

$$R = \frac{\text{variance}(x_{true})}{\text{variance}(x_{obs})}$$

then to obtain unbiased estimates we can compute $b = b^*/R$, $a = \bar{y} - b x_{obs}$. Thus, the general effect of adjusting for measurement error here is to increase the strength of the observed relationship.

correlation equal to one), then the addition of random measurement error to these would lead to the same *mathematical* result, namely that the resulting correlation would be less than perfect, and this presumably accounts for the confusion with the term "regression to the mean".

In this case, if we were to select a group with high (standardised) scores on one measure then the mean (standardised) score for this group on the second measure will be less than the mean score on the first measurement. In essence, that is one possible explanation for Feinstein's results, and indeed Jerrim and Vignoles claim that the existence of such measurement error is entirely consistent with no relative changes between SES groups over time. Clearly, then, it is important to examine and explain what is happening.

Reanalysing the data

In our reanalysis, the response variable (y) in the model is the second-occasion Millennium Cohort Study score (age 5), x is the first-occasion score (age 3), and we include grouped income (SES) as a further explanatory variable. Thus the aim of the analysis is to see whether, having adjusted for the first-occasion score, there remains an important relationship between the second-occasion score and SES. This formulation of the problem uses the available data more efficiently than Feinstein since it includes all SES groups rather than just the "rich" and "poor" groups. It also does not induce different initial mean scores for the SES groups by grouping the first-occasion test score.

Table 1 summarises our results. All scores have been normalised so that they have an overall standard normal distribution. Column 1 shows the results from fitting this model with a non-linear age 3 score and interactions with SES group, but without any adjustment for measurement error. We see that the difference between the extreme SES groups is 0.44 (0.232 + 0.207). Since the response is normalised, these are standard deviation units.

It is convenient to discuss the size of measurement errors in terms of the test score reliability, R. This is defined as

$$R = \frac{\text{true test score variance}}{\text{observed test score variance}}$$

where observed test score variance is the sum of true test score variance and measurement error variance.

Column 2 uses an adjustment for measurement error equivalent to a test score reliability of 0.75, and we see that the difference now is reduced to 0.26 standard deviation units, and for a test score reliability of 0.65 (column 3) it becomes just 0.13. As is unfortunately common with many tests, we do not have good estimates of the actual reliability of the age 3 score – hence the use of a range of possible values. But, in any case, a value of 0.65 would seem a realistic lower bound, since the testing literature reports few reliabilities below this value.

If we accept this, then the inference from this sensitivity analysis is that there is indeed an increasing inequality between the extreme SES groups (but not necessarily between the lowest and middle SES groups). However, the size and importance of this inequality are unclear and depend on the actual value of the unreliability. The greater the unreliability, the smaller the divergence between SES groups over time.

Concluding remarks

It is unfortunate that information on test reliability is absent. Such information should ideally be provided by the constructors and suppliers of the tests. Users of the data require such information. But even without a good estimate of reliability, a sensitivity analysis over a range of values illustrates the need to treat Feinstein's original conclusions with care.

On the basis of this reanalysis it would seem perfectly reasonable to conclude that there is indeed some increasing divergence between high and low SES groups over the early years in terms of developmental achievement. Whether this is as large as 0.44 standard deviation units or a more modest 0.13 is still an open question, and an important one, although the technicalities associated with the critiques clearly pose difficulties for policy-makers.

Table 1. Age 5 reading score related to age 3 reading score and SES, for different levels of test score reliability. SES group reference category is middle 50%. Standard errors in brackets. Sample size = 10 071

Parameter	*R = 1*	*R = 0.75*	*R = 0.65*
Intercept	0.038 (0.013)	0.052 (0.013)	0.042 (0.014)
Age 3 score	0.494 (0.012)	0.712 (0.016)	0.874 (0.019)
(Age 3 score)^2	−0.037 (0.006)	−0.090 (0.011)	−0.112 (0.014)
Lower 25% SES	−0.207 (0.021)	−0.092 (0.022)	−0.006 (0.025)
Upper 25% SES	0.232 (0.021)	0.168 (0.022)	0.125 (0.024)
Age 3 × lower SES	0.054 (0.021)	0.030 (0.029)	0.024 (0.032)
Age 3 × upper SES	−0.022 (0.022)	0.0016 (0.0300)	0.021 (0.029)
Residual variance	0.668 (0.009)	0.515 (0.009)	0.431 (0.011)

Estimation by Markov chain Monte Carlo with default diffuse priors: burn-in = 500, iterations = 1000.

IPGGutenbergUKLtd/iStock/Thinkstock

In my view, this suggests that a more cautious, long-term attitude should be taken towards such research findings. Social research is a highly contested area, whether published in a "reputable" journal or as a non-peer reviewed report to a sponsor. Policy-makers would do well to promote a wide debate about any findings that appear important, where technical and interpretational issues can be debated in terms that are widely accessible, and where other relevant research can be referenced.

Since many, if not most, measurements in the social sciences have embedded measurement errors, it ought to become routine for researchers to consider the implications of such measurement unreliability for their analyses. In some cases substantive conclusions may be unaffected, but those who use the results of research, including policy-makers, need to be aware of the issues. Moreover, since procedures such as those used in the present analyses, developed by Richardson and Gilks,[4] are fairly widely available, there should be no reason for data analysts to neglect their use.

Another approach to measurement error adjustment

Since they claim to have adjusted all SES differences by allowing for measurement error, a brief comment on the method used by Jerrim and Vignoles[2] is worth making. They use an instrumental variable approach; the instrument they use is a test taken at the same time as the test of interest at age 3, and they assume that the former test score is uncorrelated with any measurement error in the test of interest. This does seem, however, a very strong assumption, especially since the tests were taken on the same day. Furthermore, for instrumental variable methods where it is likely that the instrument is uncorrelated with measurement errors in the test of interest, the method will often tend to lack statistical power. In addition, there is a substantive problem in that using the instrument as a measure of "ability" assumes that it is the same "ability" that is being measured by the test of interest; in other words, it is what is known as a parallel test. This does, however, seem questionable. For these reasons we have adapted the approach of Richardson and Gilks,[4] but we do need to emphasise that in a sensitivity analysis, using more than one estimate for the measurement error variance is important.

References

1. Feinstein, L. (2003) Inequality in the early cognitive development of British children in the 1970 cohort. *Economica*, **70**, 73–97.
2. Jerrim, J. and Vignoles, A. (2013) Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society, Series A*, **176**, 887–906.
3. Goldstein, H. and French, R. (2015) Differential educational progress and measurement error. *Longitudinal and Life Course Studies* (to appear).
4. Richardson, S. and Gilks, W. (1993) Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, **12**, 1703–1722.

Harvey Goldstein is professor of social statistics at the University of Bristol's Centre for Multilevel Modelling. He is currently joint editor of the *Journal of the Royal Statistical Society, Series A*