Do different measures of assessment performance bias or exaggerate relationships?

Dave Putwain, Edge Hill University, Ormskirk, UK

Abstract

KS4 assessment performance data is used in at least three ways in educational research. First, an indicator of overall school performance to evaluate policy initiatives/ school practices designed to raise achievement. Second, to establish differential student achievement or mobility across socio-demographic groups (e.g. gender, ethnic background, etc.). Third, to establish how psycho-educational constructs such as achievement motivation, academic self-concept and assessment anxiety may influence students’ attainment. A standard convention for educational research in the UK is convert GCSE grades into an eight-point scale (e.g. Schagen & Schagen, 2005). There is less consistency over the calculation of performance measures and six different conventions are in current usage; total sum of points, total sum of points in the eight highest scoring subjects, grade point average (GPA), the mode of highest grades achieved, the number of five A*-C grades achieved and a binomial categorisation of whether a student achieved 5 A*-C grades or not. There is a possibility that particular measures may bias or exaggerate certain relationships. In order to investigate this possibility, an existing dataset regarding the relationship between GCSE performance and self-reported assessment anxiety was reanalysed using these six measures in a series of simple regression models. This reanalysis suggested that two measures, GPA and whether a student has achieved five A*-C grades or not, produce a stronger effect than the other measures. For example, GPA indicates a highly anxious student would achieve one grade lower across all subjects, whereas the total sum of points suggests a highly anxious student would drop three grades overall. These findings indicate that individual performance measures should be treated with some degree of scepticism and that in order to establish an educationally meaningful effect, multiple measures of performance should be triangulated.
Introduction

Using academic grades as performance indicators in academic research

Measures of academic performance, such as Cognitive Ability Tests (CATs), Standardised Ability Tasks (SATs) and General Certificate of Secondary Education (GCSE) grades are used as performance indicators in at least three different types of research. First, in the evaluation of policy initiatives/school practices designed to raise student performance. In this type of research, performance measures are typically used as indicators on which the success or failure of these initiatives/practices is based. Second, in order to establish how social factors influence performance. In this type of research, performance measures are typically used to assess achievement or mobility within or between different groups (e.g. gender, ethnic background and socio-economic background). Third, in order to establish how psycho-educational factors influence performance. This type of research is typically used to establish how psycho-educational constructs such as academic emotions (e.g. anxiety or pride), academic self-concept and achievement goals may influence achievement.

There now follows a brief review of some recent literature from UK based research to provide some examples of the types of research listed above concerned with GCSE performance. This review is not intended to be exhaustive, nor has research been categorised into each using a rigid, systematic procedure. Rather the purpose is to provide a context to the paper and offer a focus on the ways in which performance measures are used in educational research.

An example of the first type of research outlined above, where performance measures are used to assess the effectiveness of a school practice is found in Ireson, Hallam & Hurley (2005). These authors used a multilevel model to establish the effects of academic grouping on GCSE achievement in English, mathematics and science while controlling for prior attainment, social disadvantage, gender and attendance. GCSE grades were coded on a scale of 0-8 (8 = A*, 0 = unclassified) and separate measures were included from English, Mathematics and Science (an average science score was used where students have taken more than one science course). A further example of this type of research was reported by Demie (2005) who set out to identify factors contributing to raising achievement in Caribbean heritage pupils in Lambeth schools. In this research performance measures involved a comparison of case schools with local authority and national schools on the percentage of students attaining 5+ at Key Stage 3 and five or more A*-C grades at GCSE.

An example of the second type of research outlined above, where performance measures are used to establish how social factors influence performance can be found in Schagen & Schagen (2005). A multilevel model was fitted to data in which GCSE performance was predicted by pupil level factors (sex and Key Stage 2 attainment), school factors (indicators of faith school, specialist school, grammar school and deprivation) and local authority level factors (the percentage of pupils attending grammar, faith and specialist schools). The measure of performance used in this research was total GCSE points score using the same system as Ireson, Hallam &
Hurley (2005) above. In a second example of this types of research, Connolly (2006) used the Youth Cohort Study of England and Wales to examine the influence of social class, ethnicity and gender on GCSE attainment via a logistic regression. The measure of performance used in this study was dichotomous: whether an individual student had attained five or more A*-C grades at GCSE or not.

An example of the third type of research outlined above, where performance is predicted from psycho-educational constructs can be found in Strand (2006) who examined whether CAT and Key Stage 2 SAT scores aged 11 could predict Key Stage 3 scores and GCSE grades using a multiple regression model. Two measures of performance were used in this research. The first involved a variant of the total number of GCSE points (using the Ireson et al. system described above), referred to as the ‘Best 8’. In this system, the total number of GCSE points are calculated from the eight (maximum) highest grades. The second measure was sum of the number of higher grades (A*-C) achieved. A second example of psycho-educational research is Putwain (2006), who examined whether the relationship between assessment-related anxiety and GCSE grade was moderated by gender, ethnic background and socio-economic background using hierarchical regression models. This research also used two measures, point score in each academic subject (using the Ireson et al. system described above), and a grade point average (GPA) calculated from the total GCSE points divided by the number of subjects taken.

**Which measure is best?**

It is evident from this brief literature review, that different conventions exist for the measurement of academic performance in the UK. Some of the authors whose research was briefly described, offer a rationale for their choices. For instance, Schagen & Schagen (2005) suggest that binary indicators (such as whether a student has achieved five or more A*-C grades are “poor indicators” of performance (p. 312). These authors argue that this measure limits analysis to logistic models, is influenced by borderline pupils (although it is not specified in their article, presumably it is those students at D/C boundary) and does not account for changes at the top and bottom of the grade spectrum. In contrast, Connolly (2006) states in respect to the binary measure of five or more A*-C grades:

“This tends to be the key and most well-known measure of educational performance at this level and continues to provide the basis for comparisons in terms of gender as well as ethnicity and social class.”

(p. 8-9)

Connolly (ibid.) reports a reanalysis of data using total GCSE points, and claims:

“Interesting, while minor discrepancies emerged, the key findings remained the same whatever measure was used.” (p. 9.)

As these additional findings are not reported, we are left with little option than to take the word of Connolly at face value. Despite Connolly’s assertion that key findings
remained unchanged in this example, the possibility remains that certain measures may bias, exaggerate, or strengthen/ weaken the magnitude of an effect or relationship. For example, the measure of examination performance used in Putwain (in press), GPA, only takes into account the average GCSE grade of a student and not the number of subjects taken. For instance, two students could have the same GPA of 5.0 (i.e. an average grade C) but, in principle, achieved from a different number of subjects. It could be argued that a student with a GPA of 5.0 achieved from nine subjects has demonstrated greater achievement than a student with a GPA achieved from only five subjects, yet this measure of performance does not a reflection of this achievement.

In order to investigate the possibility that different performance measures may influence the magnitude of an effect, Putwain’s (2006) dataset was reanalysed using six different measures: GCSE points total score, ‘Best 8’, GPA, mode grade, sum of A*-C grades and whether a student had achieved five A*-C grades or not.

Method

Design and data collection

This data was originally collected to fulfil two aims. First, to establish the relationship between self-reported assessment anxiety and second, to establish whether this relationship was moderated by a number of socio-demographic variables including gender, ethnic background and socio-economic background. Assessment anxiety was measuring using the worry subscale of a 20-item instrument, the Test Anxiety Inventory, developed and validated by Spielberger (1980). This measure has been used extensively with high school students in the United States and the factorial validity well established (Benson, Moulin-Julian, Schwarzer, Siepp & El-Zahhar, 1992; Anderson & Sauser, 1995). Worry score is based on eight items using a 4-point Likert type response format (1 = Almost never, 2 = almost always) producing an observed range of 8-32. GCSE grades were collected for 1154 mixed ability Key Stage 4 students drawn from six secondary schools in the UK in all subjects. Grades were converted to a numerical value by using the Ireseon et al. system described above.

Analytic strategy

As this dataset contains the GCSE grades in each individual academic subject for each participant, it is possible to generate a series of different overall measures of GCSE performance. Total GCSE points score was calculated from the sum of individual grades, the ‘Best 8’ calculated from the sum of the highest eight individual grades, GPA calculated by dividing the total GCSE points score by the number of subjects entered for and the mode grade calculated by taking the most frequently occurring grade. In multi-modal cases the mean of the different modes was calculated. The sum of A*-C grades provides the total points sum from the higher grades only and a
binomial variable recoded to establish whether a student had achieved five higher (A*-C) grades or not.

In order to establish whether differential effects are reported for each performance measure a series of simple regression analyses were performed using worry score as the explanatory variable and each performance measure as the outcome variable. All, but the final performance measure are continuous and so were calculated using an OLS regression. The final measure, whether a student has achieved five higher (A*-C) grades or not is binomial, hence a logistic regression was calculated.

Findings

Overall model fit statistics and regression coefficients for the different performance measures are reported in table 1 below. Worry scores were highly significant predictors of all six measures of GCSE performance, however the overall model fit statistics (the $F$ and $R^2$ values for OLS regression, and the $\chi^2$ and $R^2$ values in the logistic regression) suggest that that size of the effect may vary for different performance measures, in particular a larger effect is reported for GPA. In order to establish how performance changes in relation to worry scores it is necessary to interpret the regression coefficients of the fitted models in greater detail.

Table 1: Regression analyses using different performance measures

<table>
<thead>
<tr>
<th>OLS Regression</th>
<th>$F$</th>
<th>$R^2$</th>
<th>$\alpha$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCSE points total</td>
<td>11.1</td>
<td>.04</td>
<td>29.7</td>
<td>-.31*</td>
</tr>
<tr>
<td>Best 8</td>
<td>10.6</td>
<td>.04</td>
<td>28.3</td>
<td>-.28*</td>
</tr>
<tr>
<td>GPA</td>
<td>60.4</td>
<td>.07</td>
<td>5.6</td>
<td>-.07*</td>
</tr>
<tr>
<td>Mode score</td>
<td>17.2</td>
<td>.05</td>
<td>4.3</td>
<td>-.05*</td>
</tr>
<tr>
<td>$\Sigma$ A*-C grades</td>
<td>22.8</td>
<td>.05</td>
<td>4.1</td>
<td>-.08*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logistic Regression*</th>
<th>$\chi^2$</th>
<th>$R^2$</th>
<th>$\alpha$</th>
<th>$e^\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 A*-C grades</td>
<td>14.1</td>
<td>.04</td>
<td>-.18</td>
<td>.96*</td>
</tr>
</tbody>
</table>

* Using Cox & Snell $R^2$ statistic
* $p<.001$

In OLS regression, the constant value ($\alpha$) represents performance if the explanatory variable (worry score) was 0. The unstandardised coefficients ($B$) represent the
change in GCSE performance for single unit increase in the predictor variable (i.e. worry score). As worry was measured on a scale of 8-32, β values are small and hard to interpret. In order to make these values meaningful in an educational sense, a larger incremental step is required. One approach is to transform assessment-related anxiety scores measured on a continuous scale into categories of high and low anxious using scores ± 1 standard deviation. The mean worry score was 17.15 (5.41) suggesting that students scoring in a range of 8-11 could be considered as low in test anxiety (median = 9.5) and students scoring 23-32 as high in test anxiety (median = 27). The difference between the median points of the low and high test anxiety groups in 17.5 increments. The β values in the OLS regression can be multiplied by this value to predict how each performance measure would change between a low and high self-reported worry.

GCSE points total, would be reduced by 3.3 points (grades). In educational terms this would be the equivalent of attaining three grades lower in a single subject, perhaps a grade D in Maths rather than a grade A. Alternatively, this could mean a drop of a single grade in three different subjects. Although a drop of three points (grades) from the total GCSE points score is small, the second possibility could have a great deal more serious implications both for the individual student and the school statistics. The ‘Best 8’ points total is reduced by 2.8 points, which if rounded up to three grades would result in a similar sized effect.

GPA would be reduced by 1.05 points (grades), which if rounded down would result in a one grade decrease in average GCSE grade. As with total points and ‘Best 8’ measures, this finding could have greater educational consequences if at the C/D grade borderline that at the top or bottom of the grade scale. The constant (α) value for GPA represents the grade if worry scores were 0. As this value is 5.6 (a grade C), there appears to be a real threat that a high worry score would operate around the C/D grade borderline. The Σ A*-C grades measure may help in this respect by suggesting that students reporting high worry scores would achieve 1.3 points (grades) lower, which if rounded down indicates the effect is limited to a single A*-C grade. Mode score would be reduced by 0.5 points (grades) which does really assist interpretation here (but could in other cases).

Logistic regression requires different interpretation as the outcome variable (i.e. whether a student attains five A*-C grades or not) is not continuous but binomial. The model fit statistics can be interpreted in a similar way to OLS regression, however the $e^{\hat{\beta}}$ (or odds ratio) value represents the change in the odds of a student attaining five A*-C grades for a single unit increment increase in worry score. Put simply, the $e^{\hat{\beta}}$ value of .96 indicates that as worry scores increase by an increment of 1, the odds of a student change to 96% of its previous value (i.e. a 4% decrease). A 17.5 increment increase in worry score would equate to the odds of a student attaining five A*-C grades dropping by 70%.

In summary, these findings indicate that whichever measures of GCSE performance is used, a higher worry score predicts a lower GCSE grade. However, establishing the size of the effect in an educationally meaningful sense is difficult using one measure in isolation.
Discussion

These findings indicate that, irrespective of the measure of performance used, a higher worry score is associated with a lower GCSE grade. In itself, this is not a remarkable finding, and replicates what a great deal of previous research has demonstrated (e.g. Seipp., & Schwarzer, 1996; Zeidner, 1998), albeit in alternative educational contexts than the UK. The finding of interest here relates to how the size of the effect reported by each measure and its real-life educational consequences. Total GCSE points score and ‘Best 8’ measures show a drop of three GCSE grades, but is difficult to make sense of in real terms because we don’t know where in the grade spectrum this drop is occurring. GPA indicates that a drop of one grade could possibly be occurring around the C/D borderline. The Σ A*-C grades helps to clarify the extent of this effect by indicating that is just a single A*-C grade. The binomial measure of whether a student has attained 5 A*-C grades or not helps to further clarify this effect by indicating that the odds of a student attaining five A*-C grades dropping by 70%.

The point I am trying to make here is that any measure taken in isolation would provide an effect that was difficult to interpret in an educationally meaningful sense. For example, the binomial measure when taken in isolation, could have a much greater impact that some of the other measures. Imagine the sound-bite that could be offered of this research “Stressed students 70% less likely to attain five A*-C GCSEs”. A similar argument could be constructed on the basis on the GPA measure. Such rhetoric sounds much more impressive that the conclusion that could be written of the Σ A*-C grades measure; “Stressed students gain one fewer A*-C grade”. The sound conclusion would be that without access to different types of measures, it is possible that the same data could be reported in such a way as to create a much larger sounding effect than is actually present.

Returning to the arguments of Connolly (2006) and Schagen & Schagen (2005) regarding different measures, outlined above, rather than viewing one measure as superior to another, I would argue that their mode of calculation and the subsequent analysis they produce offer different types of effects, each of which may offer distinct benefits. The total points score favoured by Schagen & Schagen (2005) would not indicate in this research the educationally critical consequence of whether the effect was taking place at the C/D borderline. The binomial method favoured by Connolly (2006) risks overstating the effect. It is only when these and other measures are combined that it is possible to make proper sense of the data.

As researchers we should explore the opportunities offered by constructing measures of performance such as GCSE in multiple ways. The reporting of statistical outcomes using different measures in full detail may be difficult given the word limits imposed by journals, and furthermore may not be relevant or indeed possible for certain analyses. Where they are however, the next best available choice would be to follow Connolly’s lead and report that re-analyses were made using alternative measures, however I would argue that we need to go further than Connolly and report, albeit briefly, how these differences might have influenced conclusions.
Correspondence

Dr. David W. Putwain
Edge Hill University
Social and Psychological Sciences,
St. Helen’s Road
Ormskirk
Lancashire
L39 4QP

putwaind@edgehill.ac.uk

References


