Analysis of bias in the CEM11+ test for Buckinghamshire

April 2018

Executive Summary

This report presents an analysis of the extent to which there is evidence of bias in the CEM11+ tests that were used for selection to Buckinghamshire grammar schools, for entry in 2014, 2015 and 2016. The analysis focusses on the pupils (approx. 4,500 each year) who attended primary schools in Buckinghamshire, for whom a range of other assessment and demographic data are available. We focus on differences by gender, socio-economic status, language status and ethnicity. The paper presents results from analysis of the 2016 cohort; results from the other years have been analysed but are presented only when the pattern of results is different in different years.

This paper presents a range of descriptive statistics for the datasets in question, and analyses group differences by free school meals eligibility (FSME), local deprivation (IDACI), English language status (EAL) and ethnicity. A range of multiple regression models have been fitted to explore the combined effect of these factors on test scores. Finally, we look at the 'predictive' relationship from CEM11+ to KS2 for different subgroups. In keeping with a wide body of research literature and tradition (eg Reynolds and Suzuki, 2014) we take evidence of differences in the relationship between measures for different subgroups as a marker for bias.

A summary of the analysis conducted is that we find no evidence of bias in the CEM11+ for any subgroup, whether by ethnicity, EAL status, gender or socioeconomic status. Any differences in the average scores of different subgroups on the CEM11+ are consistent with those subgroups' performance at KS2 and what would be expected from knowledge of their other characteristics.

Contents

Executive Summary **1**

Introduction **3**

Defining 'bias' 3

Method **4**

Data and samples 4

Analysis approach 5

Results **6**

Descriptive statistics 6

Group Differences 9

Raw differences 9

Multiple regression 14

Differential predictive validity 17

FSME vs non-FSME 17

EAL vs non-EAL 20

Conclusions **22**

References **23**

Introduction

The aim of this research is to investigate the extent to which there is evidence that the CEM11+ test is biased against any subgroups of the population. The analysis focusses on the pupils (approx. 4,500 each year for 2014, 2015 and 2016) who attended primary schools in Buckinghamshire, for whom a range of other assessment and demographic data are available. We focus on differences by gender, socio-economic status, language status and ethnicity. The paper presents results from analysis of the 2016 cohort; results from the other years have been analysed but are presented only when the pattern of results is different in different years.

This report is in part intended as a response to specific accusations of bias, for example, the claim from Local Equal Excellent that "Buckinghamshire's 11+ test contains clear and substantial bias against children from certain ethnic groups, in particular children of Pakistani and Black Caribbean heritage"[1]. This claim rests on an incorrect understanding of bias. Group differences do not, of themselves, necessarily constitute evidence of test bias, especially if differences on that test are smaller than on any other available measure. Moreover, and despite a good deal of confusion in the literature (Brown et al., 1999), bias is not the same as unfairness, although they are related. CEM's position has always been that as long as selective (grammar) schools are part of the mix, a research-driven organisation that can use its assessment expertise to reduce the selection bias against disadvantaged groups should be willing to provide that service for any selective schools that want to minimise unfairness in their selection processes.

Defining 'bias'

The literature on test bias is extensive and full of controversy. Much of it comes from the US, where racial inequality has been the focus. Inevitably, such discussions are highly charged and exist in a political context. The fact that inequality and injustice exist is not in question: of course they do. The question here is whether there is evidence that a specific test contributes to injustice.

Two key points emerge from the research literature on test bias that are worth emphasising here. The first is that group differences alone are not necessarily evidence of bias. For example, we would not say that a reading test is biased because people who cannot read score lower on it. Similarly, the fact that children whose ethnicity is identified as Pakistani score lower on the CEM 11+ and have a lower pass rate than others is no more evidence that the test is biased against them than the fact that children identified as Indian outperform the majority is evidence the test is biased in their favour. Both facts are true in this case; neither constitutes evidence of bias, *per se*.

Second, it is clear that 'armchair' inspection of tests provides a poor guide to actual bias. Numerous analyses have shown that a range of plausible claims made about the risk of certain items or item types disadvantaging particular groups often prove to be untrue when empirical data are examined (Brown et al, 1999; Reynolds and Suzuki, 2014).

Practical approaches to defining and identifying test bias generally focus on two methods: Differential Item Functioning (DIF) and Differential Predictive Validity (DPV) (Cleary, 1968; Brown et al, 1999; Camilli, 2006; Linn, 1978, 1993; AERA, APA, NCME, 1999/2014; Young, 2001; Reynolds and Suzuki, 2014). The former (DIF) is appropriate only when we have item-level data and wish to

[1] http://localequalexcellent.org.uk/index.php/2016/06/13/racial-bias-11-plus/

compare relative bias for different items or sections within a test. Here, our interest is in the overall test score as a basis for a selection decision, so the DPV approach is more relevant.

Method

Data and samples

The data comprised biographical details and test results from pupils entering into Year 7 in one local authority in England. Biographical details included gender, age, ethnicity, school (anonymized), English as an additional language (EAL), free school meal entitlement (FSME) and postcode. By matching postcode with census data we can obtain an Income Deprivation Affecting Children Index (IDACI) score. The test results were their CEM11+ assessment taken at the start of Year 6, their prior Key Stage 1 (KS1) assessment at the end of Year 2 and their subsequent Key Stage 2 (KS2) at the end of Year 6.

Much of the data required for these analyses was unavailable for children that did not attend state schools within the local authority. The dataset was provided by Buckinghamshire County Council and was therefore limited to pupils within that local authority area. Pupils in independent schools typically do not take KS2 tests and are not included in the National Pupil Database. Although there is concern in some quarters about the number of pupils from out of county and/or those attending private schools who secure places in Buckinghamshire grammar schools, without any other information about these pupils it is impossible to say whether the CEM11+ test is contributing an increase or decrease in their numbers, relative to some other selection instrument that might be invoked. For this reason, the sample used for these analyses was limited to pupils:

- having scores for KS1, KS2 and CEM 11+ and
- having data for English as an Additional Language (EAL), Free School Meals (FSME) and Income Deprivation Affecting Children Index (IDACI) and
- attending a state primary school within the county of Buckinghamshire.

KS1 and KS2 results are captured in a level of detail that must be reduced to make analysis tractable. For both assessments, there is an aggregate measure 'Average Points Score' (APS) in the National Pupil Database that summarises each pupil's overall performance on the assessment. In the case of KS2 for 2014 and 2015, the range of these APS scores is from 13.5 to 39, but the distribution is very lumpy with odd gaps and a strong negative skew. The KS2 test changed considerably in 2016 and APS was no longer available in the same format. Instead, we have a scaled score for each component. For KS1 APS, the range is 3-21 (although the 2016 dataset contains some outliers), again with a negative skew and a sharp ceiling: the highest score achieved (21) is also the mode of the distribution.

In order to try to extract a variable for these measures that had an acceptable distribution, we created a weighted sum of age standardised component scores (weighted 0.3 reading, 0.5 maths, 0.2 GPS – grammar, punctuation and spelling; these weightings are an attempt to approximate the weightings in the CEM11+ score, though the structure of that test is quite different) and then re-standardised it to have a population mean of 100 and standard deviation of 15. This scaling gives KS2

scores the same mean and standard deviation as the CEM11+ score, hence makes comparisons clearer. The population here is the full sample of 'in county' pupils (see above).

An age-standardisation process was applied to the KS1 APS scores, to the CEM11+ total score and to its component parts (verbal, maths, non-verbal).

Analysis approach

We begin with descriptive statistics for the key variables available. Then we present a series of group differences to show how different groups (for example, comparing those eligible for FSM with others) perform on the three key attainment measures, KS1, KS2 and CEM11+.

We use multiple regression analysis to see how scores on the outcome measure (CEM11+ or KS2) change with each pupil characteristic (such as FMSE or ethnicity), while simultaneously controlling for all the other demographic variables. We run three models:

1. CEM11+, predicted from: Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status
2. KS2, predicted from: Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status
3. CEM11+, predicted from: KS2, Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status

Finally, we show how the relationships among the attainment measures change for different subgroups, and after controlling for other variables. This last set of analyses essentially presents the search for differential predictive validity (DPV) that is widely seen as the crucial test for bias. Specifically, we compare the regression of KS2 on CEM11+ for different subgroups (eg FSME vs non-FSME) and with other demographic variables included as predictors in the model. Where a linear model does not fit the relationship well, we use a higher order polynomial regression.

Results

Descriptive statistics

Counts of pupils of each gender by free school meals eligibility (FSME), English as an additional language status (EAL) and ethnicity are shown in Table 1. Ethnicity in this dataset (based on National Pupil Dataset categories) is classified into six 'Major' groups (White, Asian, Black, Chinese, Mixed, Other) plus 'Missing'. These are then subdivided into 14 'Minor' codes, as shown in Table 2, together with the numbers in each classification each year. It can be seen that the ethnic mix is broadly similar each year, though there is a significant rise in the number of Asian Indian pupils in 2015 and a smaller rise again in 2016. For each year, these numbers represent the entire population of pupils within the local authority who took the 11+ test (in Buckinghamshire, pupils take the test by default unless they choose to opt out).

	N	**FSME**	**EAL**	**Ethnic 'Major' Group**						
				White	Asian	Black	Mixed	Chinese	Other	Missing
Girls	2245	128	326	1652	376	46	133	5	7	26
Boys	2306	122	373	1684	366	62	148	11	13	22
Total	4551	250	699	3336	742	108	281	16	20	48

Table 1: Cross-tabulation by gender, FSME, EAL and ethnic 'major' group, 2016 data

			Total numbers		
Ethnic Minor Code	Description	Ethnic Major	2014	2015	2016
WBRI	White - British	White	3022	3301	3172
WOTH	White - Other	White	147	173	164
AIND	Asian - Indian	Asian	93	136	170
APKN	Asian - Pakistani	Asian	459	494	487
AOTH	Asian - Other	Asian	66	108	85
BAFR	Black - African	Black	27	45	63
BCRB	Black - Caribbean	Black	35	48	41
BOTH	Black - Other	Black	7	12	4
MWAS	Mixed - White/Asian	Mixed	80	95	90
MWBA	Mixed - White/Black African	Mixed	15	24	21
MWBC	Mixed - White/Black Caribbean	Mixed	101	101	92
MOTH	Mixed - Other	Mixed	52	63	78
CHNE	Chinese	Chinese	12	17	16
OOTH	Any other ethnic background	Other	18	23	20
Missing	Missing or refused	Missing	40	32	48
		Total	4174	4672	4551

Table 2: Ethnicity categories and counts, 2014, 2015 and 2016

Income Deprivation Affecting Children Index (IDACI) information is derived from National Census data based on the pupil's home postcode[2]. The numbers of pupils in each IDACI decile are shown in Figure 1. Nationally, each of the ten deciles contains equal numbers of children, so we can see that the population of pupils in Buckinghamshire is considerably less deprived than the national average for England (decile 1 is the most deprived, 10 the least deprived).

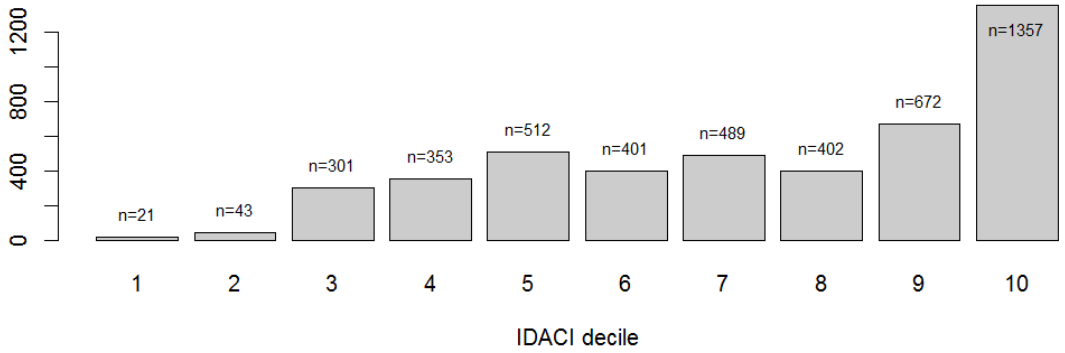

Figure 1: Counts of each IDACI decile, 2016 data (10=most advantaged 10%, 1=most disadvantaged 10% nationally)

As well as IDACI deciles, we have raw IDACI scores. These were standardised to have mean 0 and standard deviation 1.

The main variables for comparison are the three attainment measures: KS1, KS2 and CEM11+. As explained above, we created age-standardised scores from the information available for each measure. The distributions of these scores are shown in Figure 2.

[2] Obtained via the Department For Communities and Local Government website http://imd-by-postcode.opendatacommunities.org/

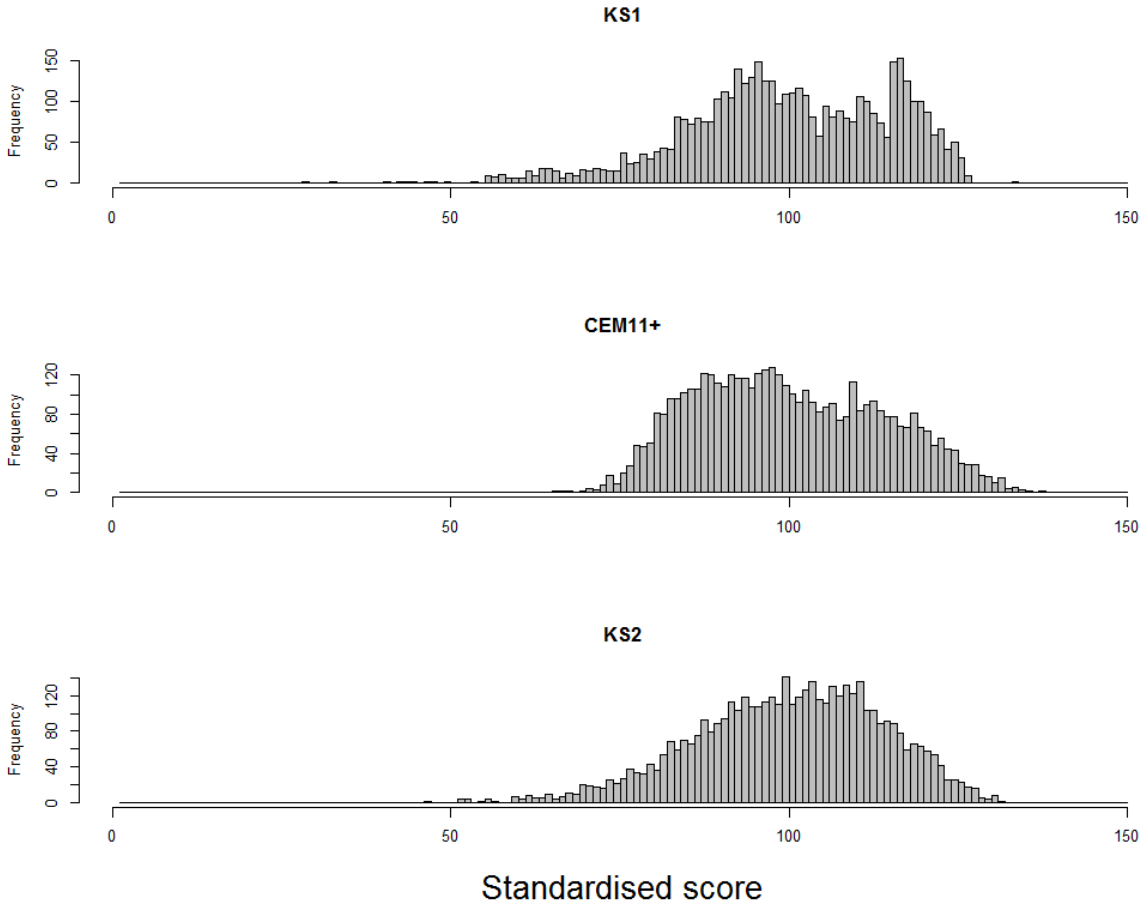

Figure 2: Distributions of age standardised scores on KS1, CEM11+ and KS2 for the full 2016 cohort

Scatter plots and correlation coefficients among these three variables are shown in Figure 3. This shows a distinct ceiling effect for KS1 and a slight curve in the relationship between CEM11+ and the other two measures.

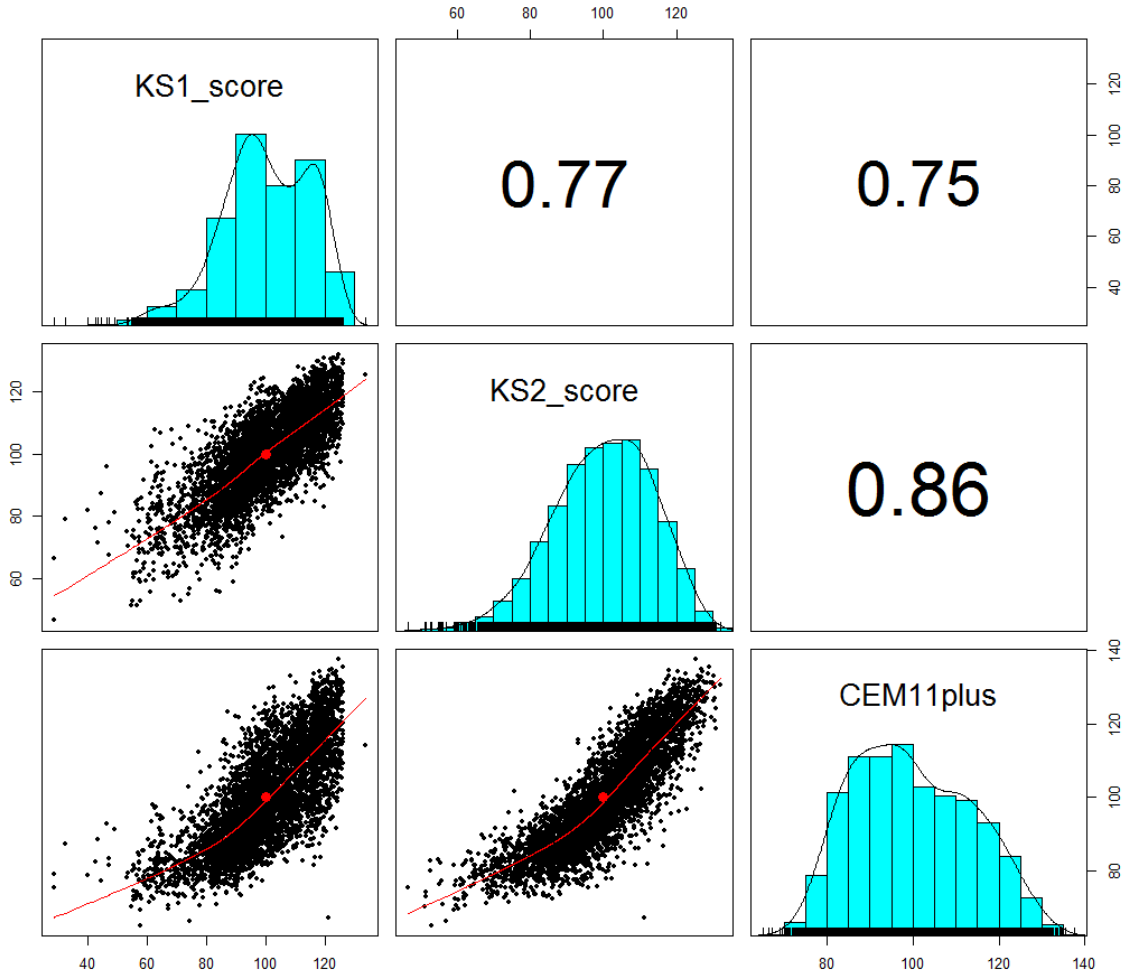

Figure 3: Pairwise scatter graphs, distributions and correlations for KS1, KS2 and CEM11+, 2016 data

Group Differences

This section presents comparisons of performance on the three attainment measures, split by gender, socio-economic status, language status and ethnicity. We first present raw differences for each group comparison separately, then adjust for interactions among them using multiple regression.

Raw differences

Group means (with 95% CIs) on each of the three outcome measures (KS1, CEM11+ and KS2) for each subgroup (by gender, FSME, EAL, Ethnic 'Major' and 'Minor' categories) are shown in Figure 4 and Figure 5.

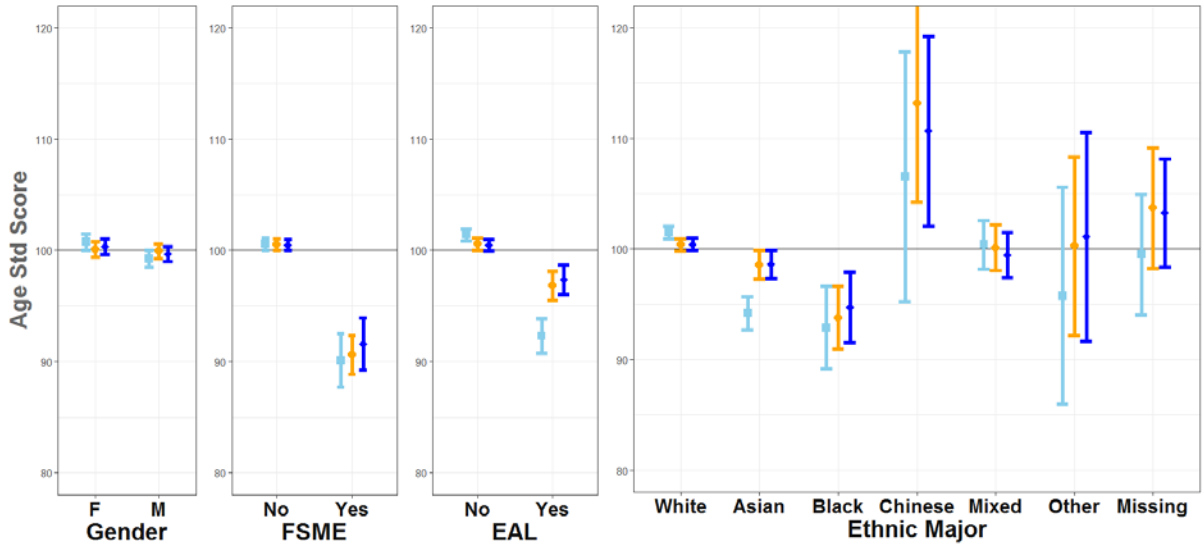

Figure 4: Group differences on KS1 (pale blue), CEM11+ (orange) and KS2 (blue) age standardised scores, by gender, FSME, EAL and ethnic 'Major' group. Group means and 95% confidence intervals, 2016 data.

As the numbers in some of these subgroups are quite small, it is important to consider the confidence intervals, which indicate the likely range of variation we would expect to see if we repeatedly drew a random sample of that size and calculated the mean. Broadly speaking, if the confidence intervals for the different measures overlap then any difference between them may be considered within the range that could easily arise by pure chance.[3]

The comparisons for 2016 data suggest that subgroup means for KS2 and CEM11+ are very close for all subgroups. KS1 scores are lower for pupils with EAL than the other two measures; they are also higher for White and lower for Asian ethnicities than KS2 and CEM11+.

[3] As the analysis applies to the full population of pupils in the 2016 cohort, there has been no random sampling and the interpretation of confidence intervals is problematic. However, the questions of interest here relate to the test not to the particular cohort available in 2016 and any inferences from the analysis are essentially claims about how other similar pupils might be expected to perform, not descriptions of the performance of these particular pupils. The particular sample of pupils analysed may be considered a representative sample, albeit not a random sample. For this reason we have included confidence intervals in the analysis in order to indicate the level of precision that should be attached to point estimates from a specific sample or subgroup. They should be interpreted as an approximate indicator of the level of precision of an estimate, not as providing any guarantee of 'confidence' in a result.

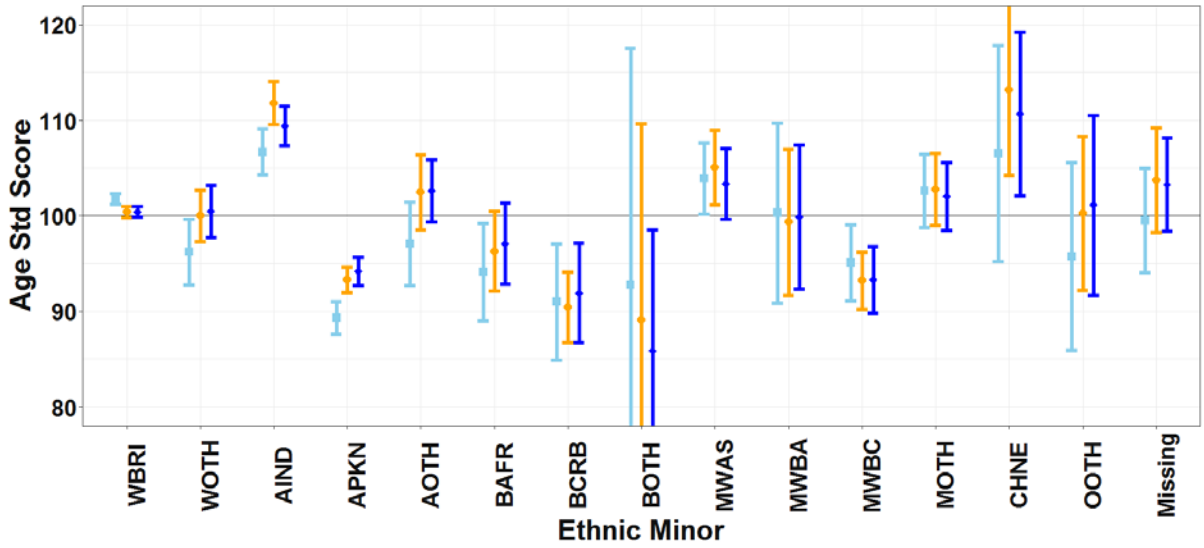

Figure 5: Group differences on KS1 (pale blue), CEM11+ (orange) and KS2 (blue) age standardised scores, by ethnic 'Minor' group. Group means and 95% confidence intervals, 2016 data.

The main comparisons of interest are between KS2 and CEM11+. KS1 is different from these two, both in the time at which it is taken (four years previously) and the nature of the assessment (teacher assessment rather than a written test). The relationship between age and score is also stronger for much younger children, which will differentially impact on the age-standardised scores. We therefore present more detailed pictures of the comparisons by gender, FSME and EAL for CEM11+ and KS2 in Figure 6 to Figure 11. These figures show histograms of the scores on each measure with the two subgroups placed one on top of the other. Also shown is a vertical line showing the mean for each subgroup (in green) and another placed one standard deviation above and one below the mean (dotted green line).

There are some small differences in the distributions of some subgroups (the overall shape of the graph) and in the size of the gap for CEM11+ and KS2 for some subgroup comparisons (these are summarised by the standardised effect size). For example, KS2 favours females slightly compared to CEM11+; the FSME gap is slightly larger on CEM11+ than on KS2. However, none of these is big enough to exceed likely chance variation: none of these differences is statistically significant.

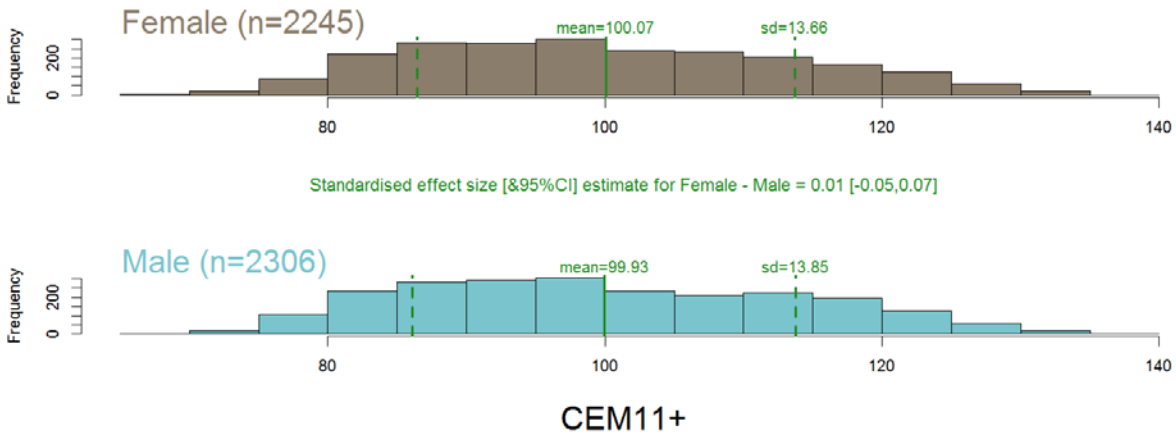

Figure 6: Distributions of CEM11+ scores by gender, with standardised mean difference effect size and confidence interval 2016 data.

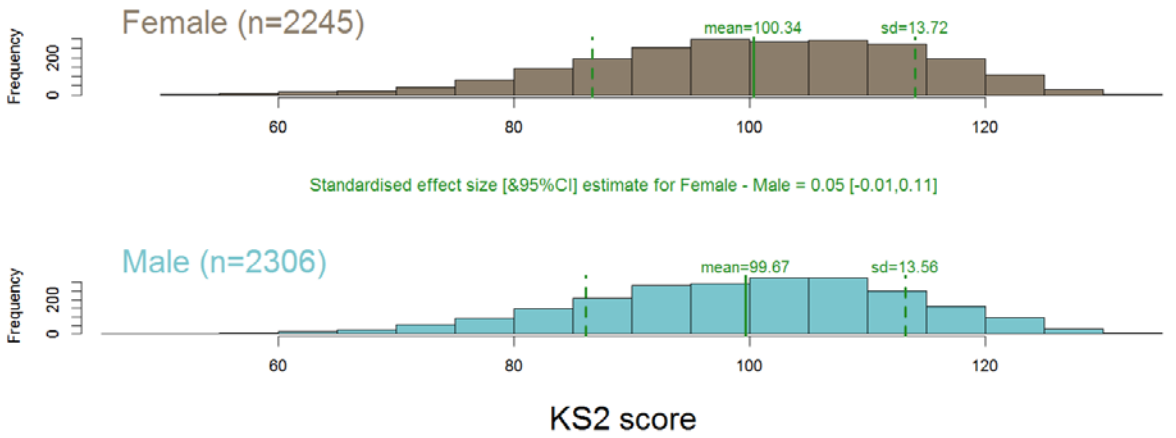

Figure 7: Distributions of KS2 scores by gender, with standardised mean difference effect size and confidence interval 2016 data.

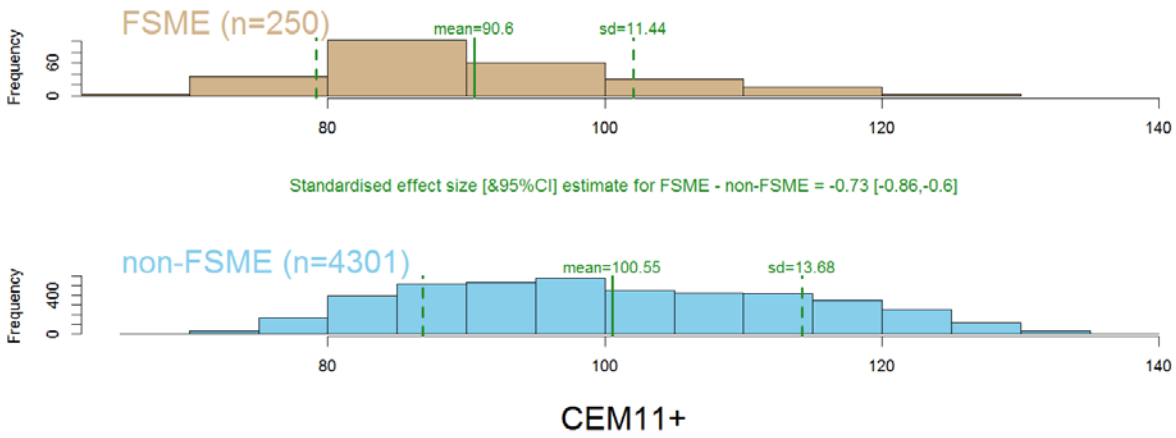

Figure 8: Distributions of CEM11+ scores by FSME, with standardised mean difference effect size and confidence interval 2016 data.

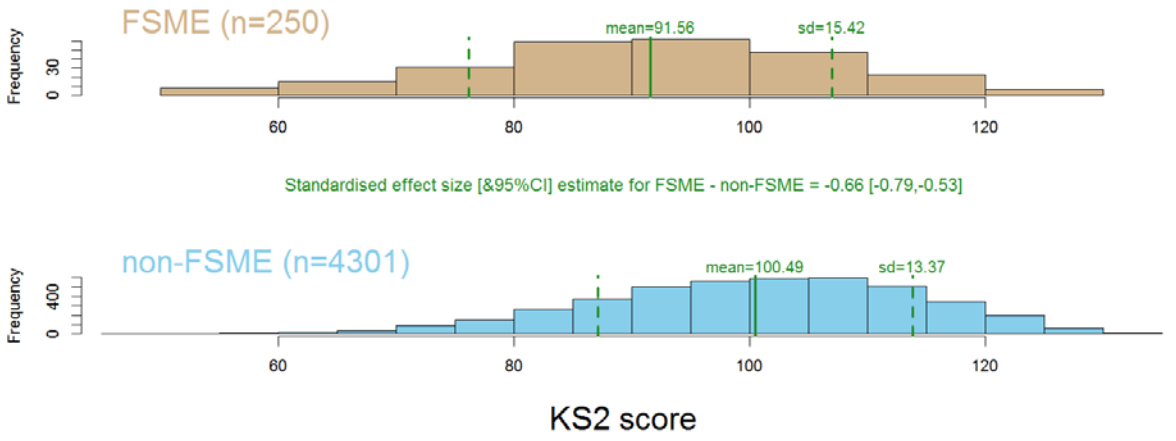

Figure 9: Distributions of KS2 scores by FSME, with standardised mean difference effect size and confidence interval 2016 data.

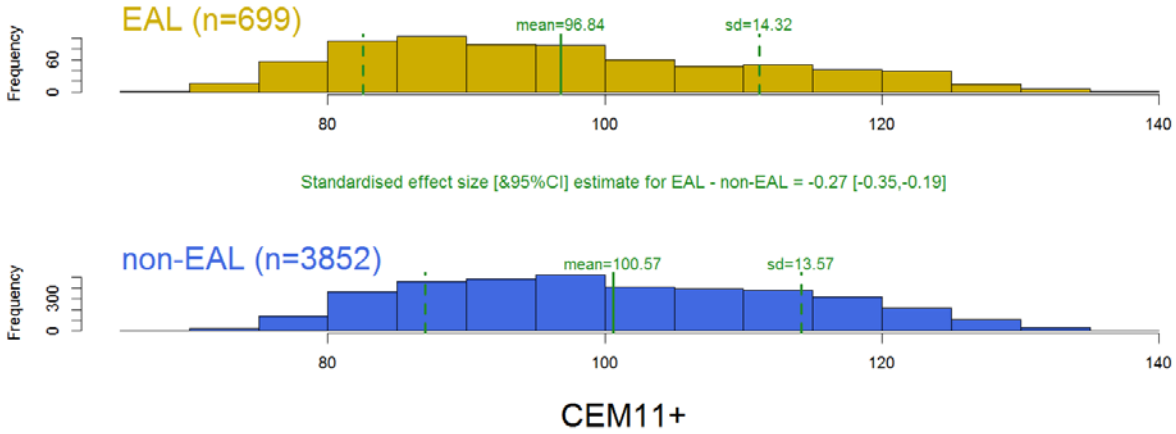

Figure 10: Distributions of CEM11+ scores by EAL status, with standardised mean difference effect size and confidence interval 2016 data.

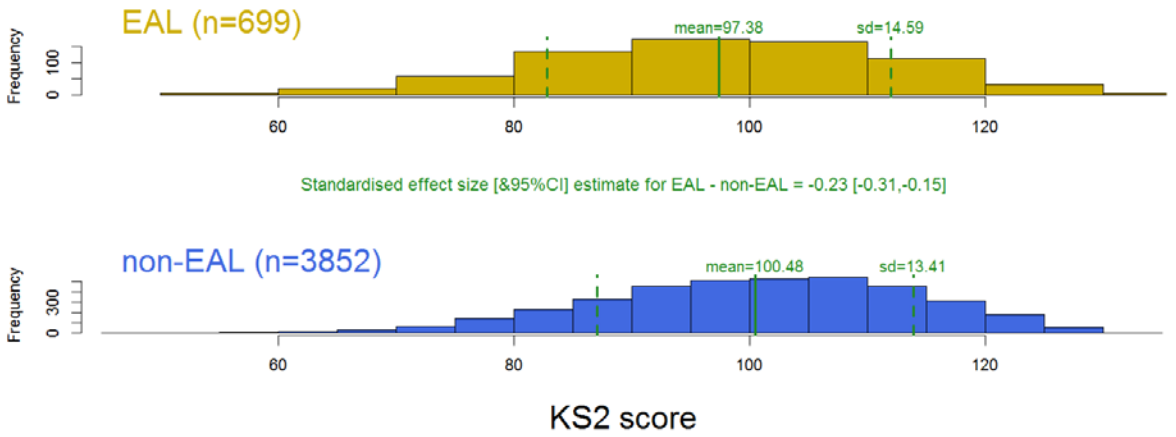

Figure 11: Distributions of KS2 scores by EAL status, with standardised mean difference effect size and confidence interval 2016 data.

Multiple regression

As stated above, we run three models:

1. CEM11+, predicted from: Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status
2. KS2, predicted from: Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status
3. CEM11+, predicted from: KS2, Gender, IDACI, FSME, EAL, and Ethnic 'Minor' status

Although IDACI and FSME are both proxies for socioeconomic status, the decision was made to retain both of them in the models. The correlation between IDACI and FSME is only 0.19 and there is no evidence of multicollinearity in any of these models. Coefficients from Models 1 and 2 are presented in Table 3 and Table 4, respectively, and the differences between these coefficients in Table 5. Model 3 coefficients are shown in Table 6.

Model 1: CEM11+ from demographic variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100.087	0.303	330.654	0.000
GenderM	-0.146	0.379	-0.386	0.700
IDACIs	-3.286	0.209	-15.716	0.000
FSMEYes	-6.182	0.850	-7.272	0.000
EALYes	-2.916	0.802	-3.633	0.000
Ethnic_MinorWOTH	2.384	1.101	2.165	0.030
Ethnic_MinorAIND	12.968	1.085	11.952	0.000
Ethnic_MinorAPKN	-0.676	0.892	-0.758	0.449
Ethnic_MinorAOTH	5.729	1.524	3.759	0.000
Ethnic_MinorBAFR	-0.914	1.637	-0.558	0.577
Ethnic_MinorBCRB	-5.721	2.020	-2.832	0.005
Ethnic_MinorBOTH	-6.341	6.395	-0.992	0.321
Ethnic_MinorMWAS	5.131	1.367	3.754	0.000
Ethnic_MinorMWBA	0.594	2.796	0.213	0.832
Ethnic_MinorMWBC	-3.291	1.366	-2.409	0.016
Ethnic_MinorMOTH	4.733	1.468	3.224	0.001
Ethnic_MinorCHNE	15.719	3.219	4.883	0.000
Ethnic_MinorOOTH	2.214	2.874	0.770	0.441
Ethnic_MinorMissing	5.288	1.875	2.820	0.005

Adj R-sq = 0.139

Table 3: Regression coefficients from: CEM11plus ~ Gender + IDACIs + FSME + EAL + Ethnic_Minor; 2016 data

Model 2: KS2 from demographic variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100.489	0.307	326.902	0.000
GenderM	-0.681	0.385	-1.767	0.077
IDACIs	-2.591	0.212	-12.203	0.000
FSMEYes	-5.864	0.863	-6.792	0.000
EALYes	-2.156	0.815	-2.646	0.008
Ethnic_MinorWOTH	2.232	1.118	1.996	0.046
Ethnic_MinorAIND	10.082	1.102	9.149	0.000
Ethnic_MinorAPKN	-1.215	0.906	-1.341	0.180
Ethnic_MinorAOTH	5.014	1.548	3.239	0.001
Ethnic_MinorBAFR	-0.781	1.663	-0.469	0.639
Ethnic_MinorBCRB	-4.997	2.052	-2.435	0.015
Ethnic_MinorBOTH	-10.727	6.494	-1.652	0.099

	Estimate	Std. Error	t value	Pr(>|t|)
Ethnic_MinorMWAS	3.312	1.388	2.386	0.017
Ethnic_MinorMWBA	0.843	2.839	0.297	0.767
Ethnic_MinorMWBC	-3.897	1.387	-2.809	0.005
Ethnic_MinorMOTH	3.588	1.491	2.406	0.016
Ethnic_MinorCHNE	12.574	3.269	3.846	0.000
Ethnic_MinorOOTH	2.711	2.919	0.929	0.353
Ethnic_MinorMissing	4.344	1.904	2.281	0.023

Adj R-sq = 0.097

Table 4: Regression coefficients from: KS2_score ~ Gender + IDACIs + FSME + EAL + Ethnic_Minor; 2016 data

Differences between Model 1 and Model 2

We can compare the regression coefficients for each variable between these two models to see how the impact of each demographic variable on the outcome (CEM11+ and KS2) differs between the two. The differences and their standard errors are shown in Table 5. We can see that in almost all cases the differences are well within the likely range of chance variation. Just one of the differences (for IDACI) reaches the $p<0.05$ threshold, but given that neither model has much explanatory power ($R^2 \approx 0.1$) and that we have estimated 20 of these differences, so should apply a suitable correction for multiple comparisons which would reduce the level of statistical significance, this does not seem significant. Overall, therefore these models suggest no significant differences in the relationships between each of our outcomes and the demographic variables available.

	diff	SEdiff	t	p
(Intercept)	-0.402	0.431	-0.932	0.351
GenderM	0.534	0.540	0.989	0.323
IDACIs	-0.695	0.298	-2.332	0.020
FSMEYes	-0.318	1.211	-0.263	0.793
EALYes	-0.759	1.144	-0.664	0.507
Ethnic_MinorWOTH	0.151	1.569	0.096	0.923
Ethnic_MinorAIND	2.886	1.546	1.866	0.062
Ethnic_MinorAPKN	0.539	1.271	0.424	0.672
Ethnic_MinorAOTH	0.715	2.172	0.329	0.742
Ethnic_MinorBAFR	-0.133	2.333	-0.057	0.954
Ethnic_MinorBCRB	-0.725	2.879	-0.252	0.801
Ethnic_MinorBOTH	4.386	9.114	0.481	0.630
Ethnic_MinorMWAS	1.819	1.948	0.934	0.350
Ethnic_MinorMWBA	-0.249	3.985	-0.062	0.950
Ethnic_MinorMWBC	0.606	1.947	0.311	0.756
Ethnic_MinorMOTH	1.145	2.093	0.547	0.584
Ethnic_MinorCHNE	3.146	4.588	0.686	0.493
Ethnic_MinorOOTH	-0.497	4.096	-0.121	0.903
Ethnic_MinorMissing	0.944	2.672	0.353	0.724

Table 5: Differences between regression coefficients from Model 1 and Model 2, with standard errors for the difference; 2016 data.

Model 3: CEM11+, including KS2 and other variables as predictors

Model 3 includes KS2 as a predictor, alongside the other predictors in Model 1. Coefficients are shown in Table 6. The proportion of variance explained (R^2) increases from 0.14 to 0.75, reflecting the high correlation (r=0.86) between KS2 and CEM11+ scores.

Coefficients in this model can be interpreted as the amount (in standardised score units, where SD=15) by which CEM11+ scores change depending on pupils' demographic characteristics. A number of these coefficients are large enough to exceed the variation that would typically be expected by chance, though the size of the difference is generally very small. For example, males achieve 0.4 scale points higher on CEM11+ than would be expected from knowledge of their KS2 score and other characteristics. Although this is a tiny difference, it appears in line with the interpretation of Figure 6 and Figure 7 that KS2 favours females slightly, while CEM11+ does not.

Those eligible for free school meals or with high IDACI scores (ie living in more deprived areas) each score about 1 point lower on CEM11+ than their KS2 score and other characteristics would suggest, as do those with English as an additional language. Pupils of Asian Indian and Chinese ethnicities have an advantage on CEM11+ (relative to the majority White British group), and by an amount that is slightly larger: of the order of 5 points. No individual ethnic group has a significant negative variation. Results from 2014 and 2015 data show exactly the same patterns and in some cases slightly larger coefficients, possibly a result of the lower overall correlation between KS2 and CEM11+ those years, but always in the same directions.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.710	0.808	20.674	0.000
KS2_score	0.830	0.008	105.321	0.000
GenderM	0.418	0.204	2.048	0.041
IDACIs	-1.136	0.114	-9.928	0.000
FSMEYes	-1.317	0.460	-2.861	0.004
EALYes	-1.126	0.433	-2.604	0.009
Ethnic_MinorWOTH	0.532	0.593	0.896	0.370
Ethnic_MinorAIND	4.603	0.590	7.805	0.000
Ethnic_MinorAPKN	0.332	0.481	0.691	0.490
Ethnic_MinorAOTH	1.569	0.822	1.909	0.056
Ethnic_MinorBAFR	-0.266	0.882	-0.302	0.763
Ethnic_MinorBCRB	-1.575	1.089	-1.447	0.148
Ethnic_MinorBOTH	2.559	3.445	0.743	0.458
Ethnic_MinorMWAS	2.383	0.737	3.235	0.001
Ethnic_MinorMWBA	-0.105	1.506	-0.070	0.944
Ethnic_MinorMWBC	-0.058	0.736	-0.078	0.938
Ethnic_MinorMOTH	1.756	0.791	2.220	0.026
Ethnic_MinorCHNE	5.287	1.737	3.044	0.002
Ethnic_MinorOOTH	-0.036	1.548	-0.023	0.982
Ethnic_MinorMissing	1.684	1.010	1.666	0.096

Adj R-sq = 0.750

Table 6: Regression coefficients from: CEM11+ ~ KS2_score + Gender + IDACIs + FSME + EAL + Ethnic Minor; 2016 data

The implications of these figures are not simple to interpret, but there is certainly no evidence to support the claim by Local Equal Excellent that the CEM 11+ test is biased against pupils identified as either Asian Pakistani or Black Caribbean. Performance of both these groups, and indeed most others, is wholly in line with their performance on the only other measure we have (KS2). Only two

ethnic groups (Asian Indian and Chinese) differ from the rest by an amount that is large enough to be both substantively important and unlikely to arise by pure chance, and both are in the positive direction: CEM 11+ advantages them relative to the majority White British group, by comparison with KS2.

Differential predictive validity

The final analysis compares the relationship between a measure to be validated (in this case CEM11+) and a criterion measure (an ideal measure of the intended construct; although KS2 is in many ways far from ideal, it provides a comparison and hence is used as the criterion measure here). We fit separate regression models for pupils in the subgroups being compared. Given previous results, we compare FSME with non-FSME and EAL with non-EAL.

FSME vs non-FSME

Regression coefficients for KS2 score as the outcome are shown for FSME pupils (Table 7) and non-FSME (Table 8). Scatter plots for KS2 against CEM11+, with the regression lines estimated in the two models, for FSME and non-FSME pupils, shown in Figure 12.

We can see that the gradient for FSME pupils (1.03) is slightly steeper than for non-FSME (0.81). This appears to suggest that proportionately fewer FSME pupils are scoring in the top range on the CEM11+ test than on KS2, with the implication that relatively fewer FSME pupils will pass a selection threshold on CEM11+ than would be the case if they were selected on KS2 scores. In general, if one subgroup has a steeper line that another in a prediction from a new measure to an established criterion measure, we would interpret this as evidence of bias in the new measure against that subgroup. However, in this case we can also see that the relationship between KS2 and CEM11+ is curvilinear and is not well captured by a straight line, so it may be that the gradient of a linear fit is not a good indicator of bias.

One simple way to investigate this is to add a squared term (CEM11+^{2}) to the model. This allows the fit line for the relationship between KS2 and CEM11+ to follow a quadratic function, and hence curve downwards. The addition of this squared term improves the fit substantially (adjusted R^2 values of 0.65 and 0.75 for FSME and non-FSME respectively) and also brings the two curves into closer alignment (see Figure 14). The curvature of the fit line for FSME pupils is greater than for non-FSME, and at the top end of scores (where the 11+ selection decision is made) the curve is even slightly lower for FSME than non-FSME, which might be interpreted as suggesting bias in the other direction (that CEM11+ favours FSME pupils), though the numbers at this end are too small to infer that confidently. However, with these models there is no evidence to suggest DPV or bias against FSME pupils who took the CEM11+ test.

FSME

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.128	5.846	-0.364	0.716
CEM11plus	1.030	0.061	16.935	0.000
IDACI	5.000	8.085	0.618	0.537
EALYes	-0.160	2.560	-0.063	0.950
Ethnic_MajorAsian	-1.913	2.570	-0.744	0.457
Ethnic_MajorBlack	1.494	3.135	0.476	0.634
Ethnic_MajorMixed	-0.916	2.787	-0.329	0.743
Ethnic_MajorOther	11.659	9.929	1.174	0.241
Ethnic_MajorMissing	-3.200	10.201	-0.314	0.754

Adj R-sq = 0.544; F(8, 242) =38.25

Table 7: Regression coefficients for KS2 score as the outcome, for FSME pupils; 2016 data

non-FSME

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.061	0.921	20.688	0.000
CEM11plus	0.812	0.009	94.888	0.000
IDACI	4.770	1.612	2.958	0.003
EALYes	0.224	0.433	0.516	0.606
Ethnic_MajorAsian	-3.278	0.436	-7.521	0.000
Ethnic_MajorBlack	-3.175	0.784	-4.051	0.000
Ethnic_MajorChinese	0.254	1.819	0.139	0.889
Ethnic_MajorMixed	-1.639	0.471	-3.482	0.001
Ethnic_MajorOther	0.651	1.595	0.408	0.683
Ethnic_MajorMissing	-1.282	1.337	-0.958	0.338

Adj R-sq = 0.695; F(9, 4411) =1118.97

Table 8: Regression coefficients for KS2 score as the outcome, for non-FSME pupils; 2016 data

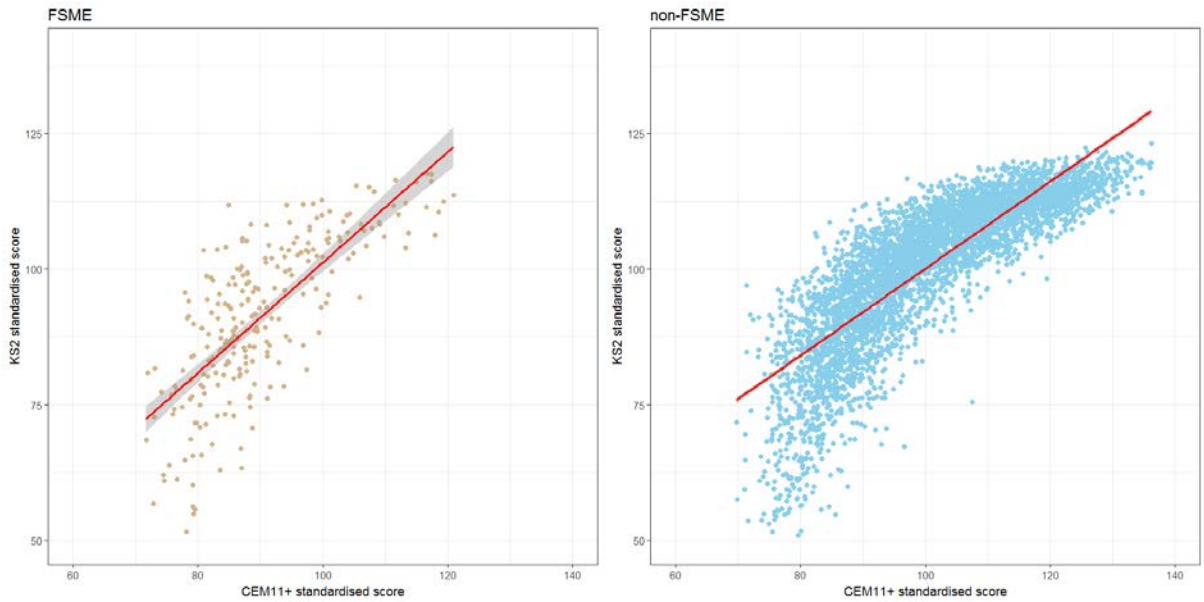

Figure 12: Scatter plots and regression lines for KS2 against CEM11+ for FSME and non-FSME pupils; 2016 data

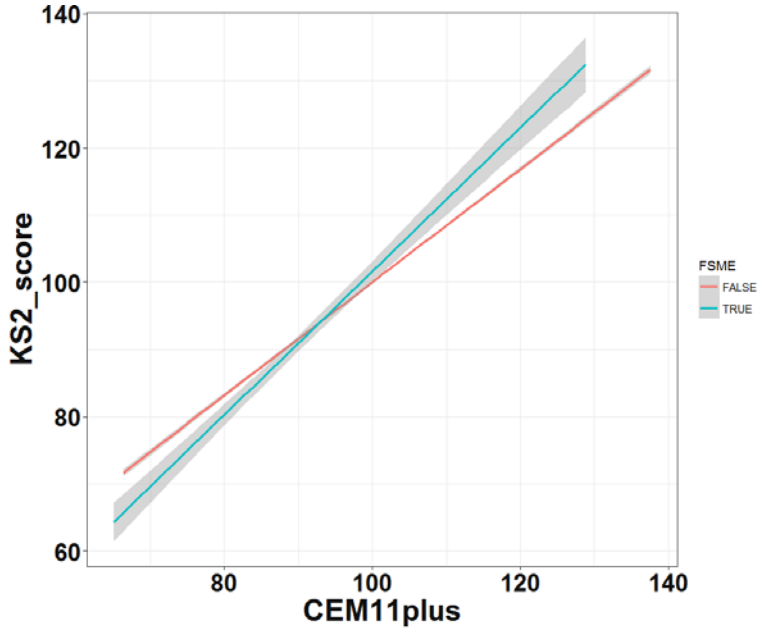

Figure 13: Regression segments for KS2 against CEM11+ for FSME and non-FSME

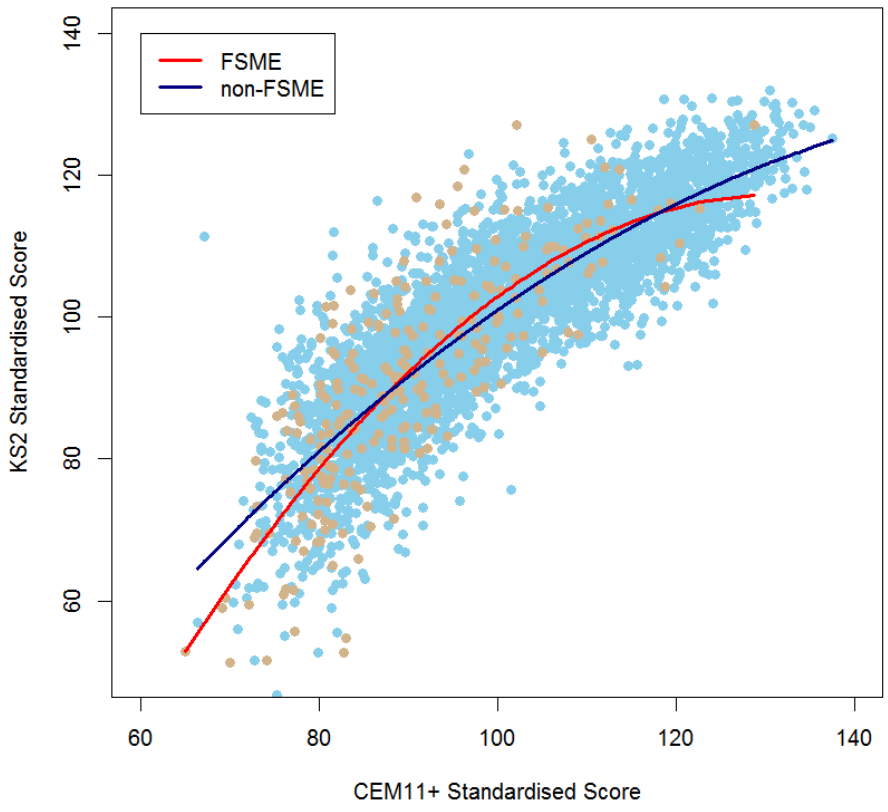

Figure 14: Scatter plots and fit lines from quadratic model for FSME and non-FSME pupils

EAL vs non-EAL

Regression coefficients for KS2 score as the outcome are shown for EAL pupils (Table 9) and non-EAL (Table 10). Scatter plots for KS2 against CEM11+, with the regression lines estimated in the two models, for EAL and non-EAL pupils are shown in Figure 15.

For EAL vs non-EAL, the gradients of the CEM11+ are much closer: 0.89 for EAL, 0.85 for non-EAL. This suggests there is little difference in the differential predictive validity of CEM11+ by EAL status.

EAL

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.852	2.446	4.438	0.000
CEM11plus	0.887	0.022	41.259	0.000
IDACI	12.000	3.707	3.237	0.001
FSMEYes	0.045	1.047	0.043	0.965
Ethnic_MajorAsian	-1.571	0.798	-1.968	0.049
Ethnic_MajorBlack	2.253	2.235	1.008	0.314
Ethnic_MajorChinese	-1.009	2.950	-0.342	0.732
Ethnic_MajorMixed	-1.627	1.967	-0.827	0.409
Ethnic_MajorOther	-5.960	3.183	-1.872	0.062
Ethnic_MajorMissing	-1.224	2.024	-0.604	0.546

Adj R-sq = 0.732; F(9, 689) =213.11

Table 9: Regression coefficients for KS2 score as the outcome, for EAL pupils; 2016 data

non-EAL

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.237	0.919	16.580	0.000
CEM11plus	0.848	0.009	98.543	0.000
IDACI	0.572	1.581	0.362	0.718
FSMEYes	-0.625	0.524	-1.192	0.233
Ethnic_MajorAsian	-0.204	0.493	-0.414	0.679
Ethnic_MajorBlack	-0.593	0.723	-0.820	0.413
Ethnic_MajorChinese	-1.665	2.297	-0.725	0.469
Ethnic_MajorMixed	-0.696	0.445	-1.565	0.118
Ethnic_MajorOther	3.737	1.840	2.031	0.042
Ethnic_MajorMissing	-0.060	1.219	-0.049	0.961

Adj R-sq = 0.738; F(9, 3842) =1207.21

Table 10: Regression coefficients for KS2 score as the outcome, for non-EAL pupils; 2016 data

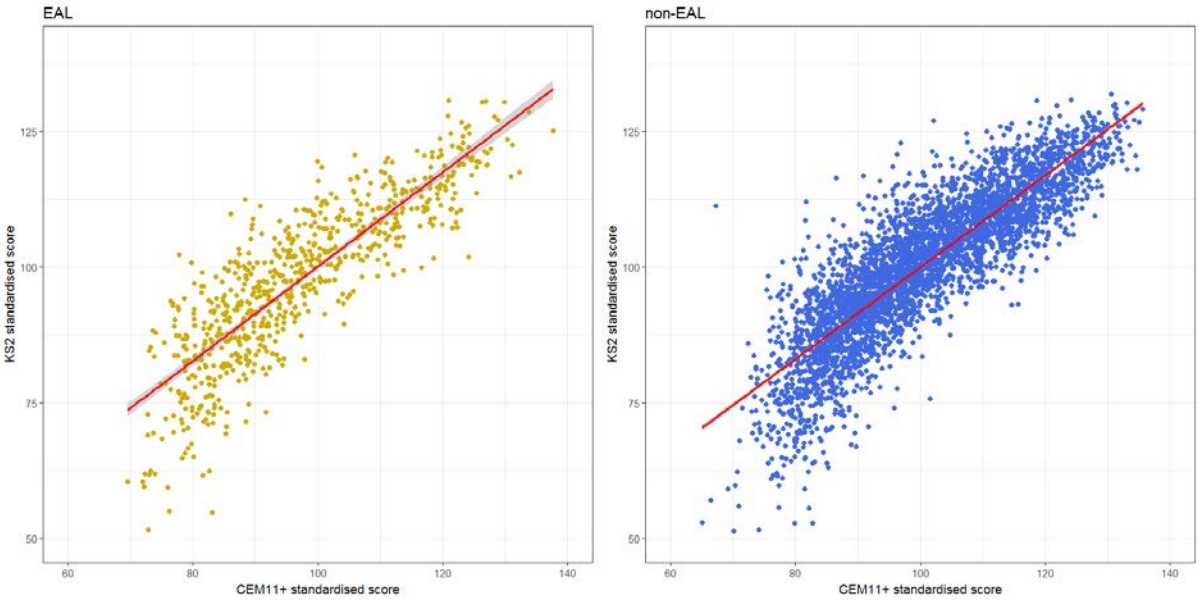

Figure 15: Scatter plots and regression lines for KS2 against CEM11+ for EAL and non-EAL pupils; 2016 data

Conclusions

This report has analysed data from pupils in primary schools in Buckinghamshire for entry into grammar schools in 2014, 2015 and 2016. The focus of presentation in the report has been the 2016 data, but in every case the other years the same analyses have been run on the other years and the results cross-checked. A range of group differences have been compared in different ways for subgroups of pupils by gender, free school meals eligibility, English as an additional language status and ethnicity.

There are some important limitations of this analysis. In principle, we cannot say definitively whether CEM11+ test is unbiased, only whether it is less (or more) biased than the KS2 test. If unequal social processes introduce unfair disadvantage for certain groups it is generally impossible for a good measure to eradicate resulting deficits in performance (Reynolds and Suzuki, 2014), even if such a course were deemed desirable. More specifically, for the current analysis to provide a meaningful comparison, it must be limited to a dataset that contains KS2 scores. This excludes candidates for Buckinghamshire grammar schools from out of County and independent schools. Without access to their KS2 scores, or some equivalent proxy, we can say nothing about how the CEM11+ test contributes bias in selection decisions.

The main measures under comparison here are the KS2 test scores and CEM11+ scores. Comparisons of raw differences across these groups show that there are substantial differences in the performance of particular subgroups on both measures. There are also some 'differential differences', where the subgroup difference is different on different measures. However, none of these differences looks to be large enough both to exceed likely chance variation and to have practical significance.

The key approach to analysis of bias employed here draws on regression analysis and in particular differential predictive validity (DPV). This analysis addresses the question of whether candidates with the same scores on the measure to be validated, and equivalent other characteristics, go on to perform in systematically different ways on a criterion measure. Ideally, a criterion measure should be an indicator of something important and valuable that is itself taken to be free from bias. In practice, we can only ever validate one measure as unbiased relative to another. The use of KS2 as the criterion measure for this analysis is certainly less than ideal: it has a number of measurement deficiencies of its own, including a range of likely biases. Nevertheless, it is all that is available currently.

A summary of the analysis conducted is that we find

- There is no evidence of bias in the CEM11+ against any minority ethnic group. It is possible that some minority ethnic groups may be advantaged by the CEM11+, relative to the majority White British group, by comparison with their KS2 performance.
- CEM11+ is broadly neutral in relation to EAL status
- CEM11+ marginally redresses a bias against males in KS2
- There is no evidence of bias in the CEM11+ in relation to socioeconomic status, whether indicated by their free school meals status or IDACI scores.
- A few subgroup differences exceed the range of likely chance variation, but all differences are small and none is substantively important.

References

AERA, APA, NCME (1999/2014). (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing) *Standards for educational and psychological testing*.

Brown, R. T.; Reynolds, C. R. and Whitaker, J. S. (1999) Bias in mental testing since 'Bias in Mental Testing'. *School Psychology Quarterly*; 1999; 14, 3, 208-238.

Camilli, G. (2006). Test fairness. In Robert L. Brennan (Ed.), *Educational Measurement*, 4th Edition pp221-256. Westport, CT: American Council on Education and Praeger Publishers.

Cleary, T.A. (1968) Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63(4), 507.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. Holland and H. Wainer (eds) *Differential item functioning*, pp349-364. Hillsdale, NJ: Lawrence Erlbaum.

Reynolds, C.R. and Suzuki, L.A. (2012) 'Bias in Psychological Assessment: An Empirical Review and Recommendations' in I.B. Weiner (Ed) *Handbook of Psychology (2nd Edition), Vol 10*. Wiley [Avail at: http://lp.wileypub.com/HandbookPsychology/SampleChapters/Volume10.pdf]

Young, J. W. (2001). Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis. Research Report No. 2001-6. College Entrance Examination Board.