RESEARCH AND ANALYSIS

Equality impact assessment: literature review

ofqual

Authors

This report was written by Ming Wei Lee and Merlin Walter, from Ofqual’s Strategy Risk and Research directorate.

Contents

Executive Summary

This paper provides a review of the literature on the nature and extent of any bias that might arise in using centre assessment grades for this summer’s cohort of GCSE, AS and A level students.

Studies of potential bias in teacher assessment suggest that differences between teacher assessment and exam assessment results can sometimes be linked to student characteristics like gender, special educational needs, ethnicity and age. However, such effects are not always seen, and when they are present, are small and inconsistent across subjects.

Teachers' grade prediction/estimation has been studied in research using UCAS data to examine the accuracy of the A level grade predictions made by teachers for their students' university admission applications and in other research using individual exam boards' data to examine the accuracy of the GCSE and A level grades that exam boards used to collect from teachers.

The same distribution of exactly accurate and over-/under-predictions and pattern of attainment-dependent prediction accuracy have been found in the two strands of research. Findings on individual variables are also broadly similar: subject has a small but unsystematic effect; gender and age have small effects which are inconsistent across subjects; centre type has a small effect which can be attributable to the correlation between centre type and attainment and attainment-dependent prediction accuracy.

There are likely some effects on prediction accuracy of ethnicity (that is, more over-prediction for some ethnic minority groups) and disadvantage (that is, more over-prediction for the more disadvantaged in general, and less over-prediction for the more disadvantaged among high attainers), but those effects have not been properly estimated.

Introduction

Following the cancellation of public exams to help fight the spread of the coronavirus (COVID-19), Ofqual has developed a process for awarding this summer's GCSEs, AS and A levels. Schools and colleges are being asked to provide centre assessment grades for their students for each subject and to rank order the students within each grade. The centre assessment grades will be standardised using a statistical model. This paper provides a review of the literature on the nature and extent of any bias that might arise in using centre assessment grades for this summer's cohort of GCSE, AS and A level students. It has three parts, focusing on:

- teacher assessment (as opposed to exams) in general
- teacher-predicted A level grades used in university admission in the UK
- teacher-estimated grades that used to be used by exam boards as one source of evidence in awarding (setting grade boundaries)

It should be noted that in assessing the nature and extent of any bias in teacher assessment, grade prediction or estimation, the research literature addresses potential bias in teachers' absolute judgments, not their relative judgments (that is, their rank ordering of students). Arguably, for our purpose, it is the potential bias in teachers' relative judgment that is more relevant but we have been unable to find research that directly scrutinises rank ordering. The extent to which any bias in absolute judgment is also present in relative judgment is not known, but any bias causing one of two equally able students to be over- or under-marked/predicted/estimated in absolute judgment can be expected to affect the ordering of the two students in relative judgment. Therefore, we consider that a review of the literature on any bias in teachers' absolute judgments of students is highly relevant in the present context.

Teacher assessment (as opposed to exams) in general

Despite literature suggesting that teachers' estimates generally correlate well with grades achieved in examinations (Dhillon, 2005) and have potentially greater validity than formal tests (Hoge & Coladarci, 1989), there is also a range of evidence that highlights issues of low reliability and potential bias in teacher assessments (reviewed in Harlen, 2005) relating to a range of student characteristics, including gender and special educational needs as well as ethnicity and age (Harlen, 2005; Fajardo, 1985; Reeves et al., 2001).

Reeves et al. (2001) compared teacher assessment with exam testing in a study of approximately six thousand Key Stage 2 pupils in LEA-maintained schools over three years (1996-1998). This study found that there were significant effects of age, gender, SEN (special educational needs) and EAL (English as an additional language) on teacher assessment scores in English, maths and science, such that older pupils (that is, those born earlier in the same school year) were rated slightly higher than younger pupils across all subjects, and male pupils were rated higher in maths and science whereas female pupils were rated higher in English. SEN pupils were rated lower than non-SEN pupils across all subjects and years, and EAL pupils were rated lower than non-EAL pupils in 1998, whilst in 1996 and 1997 the authors suggest that the number of EAL pupils may have been insufficient to detect any difference.

Although the authors report numerous significant effects of the studied characteristics on teacher assessment scores, these were often similar to the effects of those factors on exam scores. When the two assessment methods were directly compared, teacher assessment and exam scores reportedly differed significantly across gender such that teachers tended to underrate male pupils in maths and science, but to underrate female pupils in English, relative to test results. The authors of this study highlighted that, although statistically significant, these differences were small and only affected a small proportion of pupils.

The results of this comparison also showed that teacher assessment consistently and markedly underrated SEN pupils in English and science, relative to exam scores, but did so in only 1998 for mathematics. EAL did not significantly correlate with differences between teacher assessment and exam scores in science or mathematics, and only did so in 1998 for English where EAL pupils were underrated relative to their exam scores. This isolated effect was noted as being particularly strong. Of the three years and subjects studied, pupil's age only significantly affected maths in 1998, and this effect was minimal.

The authors also noted that where differences between teacher assessment and exam scores are only of one level or grade, these may represent pupils whose general performance is close to the boundary and thus may not be cause for any particular concern. Differences of more than one level were considered to be suggestive of a 'substantive gulf' between teachers' views and pupils' performance, and these only occurred in 57 cases out of 16,087 pairs of test-teacher assessments across the three years.

Marcenaro-Gutierrez & Vignoles (2015) studied a sample of 3,000 Spanish primary and secondary school pupils using a regression modelling method, and found that in maths teacher assessments tended to underrate male pupils and overrate female pupils, relative to their test scores. These differences were not observed in reading assessments. The socio-economic/demographic factors examined (immigration status, parents' education, and household cultural capital) were not significantly associated with differences between teacher assessment and test scores, although those pupils with lower attendance were underrated in teacher assessments, compared to test performance.

Spear (1984) found that scripts from secondary school students of either gender were graded higher by science teachers if they were perceived to have been written by male students, and Fajardo (1985) studied the influence of ethnicity on grading using essays (for which a 'definitive' mark had been previously determined) accompanied by 'bogus' admission forms, reporting that teachers and student-teachers tended to rate essays perceived to be written by black students higher than those whose ethnicity was not indicated.

Conversely, Baird (1998) examined the potential benefits of blind marking in the marking of A level English literature and chemistry scripts, and found no evidence to suggest that marking was biased along gender lines, and an investigation by the Scottish Examination Board (1992) found that teachers did not exhibit any observable gender or ethnicity bias in the marking of English scripts, but found that history scripts perceived to have been written by female students tended to receive higher marks than those supposedly written by male students.

Mastergeorge and Martinez (2010) conducted a study examining the judgments of active teachers on disabled and general education pupils aged 9 to 15 in the south-western United States. The authors found no systemic bias for or against pupils with disabilities in any of the subjects investigated (arts, languages and mathematics), but did observe that teachers marked pupils with disabilities less consistently than general education pupils.

Gibbons and Chevalier (2008) examined large-scale administrative datasets on students aged 11 to 16 in England from 1997 to 2004, containing details of approximately two million students, and suggested that there may be systematic differences between teacher-assessments and test performance that vary by ethnicity, gender and socioeconomic background. This study reported that ethnic minority students were rated significantly lower in English, but significantly higher in maths, relative to their test scores, whilst in science there was no significant difference observed between these assessment methods.

Male students were found to fare better on teacher assessments in English, relative to their test scores, than girls, but were underrated in mathematics and science, and older students were rated better in teacher assessments in all subjects, particularly so in science. No significant differences between teacher assessment and test scores were seen for FSM (eligible for free school meals) students compared to non-FSM students, and EAL students were significantly disadvantaged in mathematics but not in English or science.

The authors of this study observed that there was a ‘substantial negative association between prior achievement levels and the gaps between teacher and test scores’, such that lower (prior) achieving students performed relatively better in teacher assessments and that higher achieving students performed relatively better in tests, and that this effect was markedly larger than the ‘modest’ effects of social factors, ethnicity and gender.

Krkovic et al. (2014) studied a sample of 1,500 Finnish 6th grade pupils (aged 11 to 13) using multilevel analyses which accounted for between- and within-class effects. They reported that pupils’ gender influenced teachers’ assessment of their first language (Finnish) ability, and the teachers’ judgement of their potential for general academic success such that in both cases female pupils were rated higher relative to boys when controlling for test performance. Gender was not observed to affect teacher assessment of pupils’ ability in maths, and neither the teacher’s gender nor the interaction between the teacher’s gender and the pupil’s gender (same vs different) showed any significant effect on teacher assessment in any subject.

Thomas et al. (1998) used a multilevel modelling approach to examine the relationship between pupil characteristics (gender, age, SEN, EAL, and FSM) and performance in teacher assessment and standard task attainment in English, maths and science for a sample of approximately 17,000 young (aged 7) pupils in the UK. They reported broadly comparable results for both teacher assessment and standard tasks, such that although there were significant differences in pupils’ performance associated with the tested characteristics, these were similar regardless of the assessment method used. Teacher assessments were observed to ‘modestly’ widen the gap between SEN and non-SEN pupils, compared with the standard task, in all subjects, as well as for EAL pupils in English only.

In summary, there is some evidence to suggest that pupils' characteristics can influence teacher assessments. However there are also numerous studies that have found no effect, or effects that vary across subject lines. Some of the studies that have identified differences due to pupil characteristics have highlighted the small magnitude of these effects (Reeves et al., 2001; Gibbons & Chevalier, 2008) and others have identified similar or greater effects associated with prior attainment or ability (Gibbons & Chevalier, 2008; Hoge & Butcher, 1984), behaviour or attendance (Marcenaro-Gutierrez & Vignoles, 2015), and the nature of the school attended (Reeves et al, 2001).

At a glance summary: findings of the reviewed studies

	Reeves et al (2001)			Marcenaro-Gutierrez & Vignoles (2015)		Spear (1984)	Fajardo (1985)	Baird (1998)		Scottish Examination Board (1992)		Mastergeorge & Martinez (2010)			Gibbons & Chevalier (2008)			Krkovic et al (2013)			Thomas et al (1998)		
	Eng	*Math*	*Sci*	*Read*	*Math*	*Sci*		*Eng*	*Chem*	*Eng*	*Hist*	*Art*	*Lang*	*Math*	*Eng*	*Math*	*Sci*	*Lang*	*Math*	*All*	*Eng*	*Math*	*Sci*
Gender[1]	▲	▼	▼	-	▼	▲		-	-	-	▼				▲	▼	▼	▲	-	▲	-	-	-
Age[2]	-	▼	-																		-	-	-
Ethnicity[3]				-	-		▲			-	-				▼	▲	-						
SEN	▼	▼	▼									-	-	-	-	▼	-				▼	▼	▼
EAL	▼	-	-																		▼	-	-

▼ Underrated by teacher assessment

▲ Overrated by teacher assessment

- No observable difference between teacher assessment and exam performance

[1] Male students relative to female students
[2] Older students within year group
[3] Non-white students relative to white students

Teacher-predicted A level grades in university admission

In the UK, predicted A level grades are used in university admissions. How well applicants' predicted grades match their achieved grades is routinely reported in UCAS's annual end-of-cycle report (see, for example, Chapter 8 of the latest report, UCAS 2019). The accuracy of the predicted grades has also been studied as a potential source of bias in university admissions. The studies can be divided into those that examined the prediction accuracy of individual A level grades and those that looked mainly at the prediction accuracy of individual applicants' grade points. An applicant's predicted/achieved grade points is calculated by converting each predicted/achieved A-level grade into points (A*=6, A=5, B=4, C=3, D=2, E=1) and summing the points converted from the predicted/achieved grades of their best three A levels. A grade prediction for a single subject is highly likely made by a single teacher. A grade points prediction is the calculated combined result of, in all likelihood, more than one teacher's independent prediction in each subject.

In the literature, the levels of 'exact' accuracy reported range from 42% to 52% in studies of grades, and are about 16% in studies of grade points. When the criterion for accuracy is relaxed to allow +/- one grade or +/- three grade points (given that three points is equivalent to three one-grade differences), the level of accuracy is at least 85%. When predictions are not on-the-dot accurate, they are overwhelmingly more likely to be over- than under-predictions.

All the studies examined one or more background variable and documented the characteristic pattern of exactly accurate and over-/under predictions in subgroups of applicants. Although these univariate analyses give a wealth of information, they can sometimes be misleading because they examine one variable at a time and do not consider possible correlations and/or interactions between variables. Multivariate analyses which allow the effect of a variable to be examined while holding other variables constant provide more nuanced results. In the following, we summarise results of the univariate analyses and mention results of any relevant multivariate analysis.

Studies of A level grades

The studies that fall into this category are Delap (1994a), which analysed 2,977 applicants' 9,079 grades in 1991; BIS (2011), which analysed 97,268 applicants' 219,744 grades in 2009; and BIS (2013), which analysed 132,824 applicants' 177,094 grades in 2010. In addition, one section of UCAS (2016) contains an analysis of the effect of subject, using data from 2010 to 2015 on applicants predicted to achieve at least 12 grade points. Delap and UCAS carried out both univariate and multivariate analyses. The BIS reports contained only univariate analyses.

Attainment

Prediction accuracy varies by the predicted/achieved grade itself. For example, BIS (2013) reported that more predictions of grades A and E (64% and 57%) turned out to be accurate than predictions of grade C (39%). Delap's multivariate analysis, using the degree of departure of the predicted grade from the achieved grade as the dependent measure, found, for example, that achieved B grades were more likely to have been accurately predicted than achieved E grades (achieved B and E grades were over-predicted on average by 0.02 and 3.33 grades respectively).

Gender

All studies reported little or minimal difference between predictions for males and females in univariate analyses, but Delap's multivariate analysis found less difference between predicted and achieved grades (0.15 grade less over-prediction, that is, more accuracy) for females.

Socio-economic status

Both BIS studies reported that the higher the social class, the more accurate the predictions and the less over-prediction. From the 2011 study, the proportion of exactly accurate predictions was 58% and 43% for applicants from the highest ('Higher managerial') and the lowest ('Routine') group respectively, and the proportion of over-predictions was 36% and 50% respectively. The differences between socio-economic groups in likelihood of under-prediction were tiny compared to the differences in likelihood of exactly accurate prediction and over-prediction. The 2013 report cautioned that the apparent effect of social class may reflect the association between social class and attainment.

Ethnicity

Delap and BIS (2011) reported fewer exactly accurate predictions and more over-predictions for Black and Asian applicants than for White applicants and, only in BIS 2011, Mixed race applicants. In BIS (2011), the proportion of exactly accurate predictions was 53%, 53%, 47% and 39% for White, Mixed race, Asian and Black applicants respectively, and the proportion of over-predictions was 40%, 41%, 47% and 54% respectively. No ethnic group was more or less prone to under-prediction. BIS (2013) reported the same pattern of the least exactly accurate predictions and the most over-predictions for all Black subgroups combined, but much variation among the Asian and Mixed race subgroups: Asian Bangladeshi applicants showed a similar pattern to the Black subgroups combined, while Asian Chinese and Mixed White and Asian applicants, like White applicants, had a high level of exactly accurate prediction and a low level of over-prediction.

Delap reported that the effect of ethnicity found in his univariate analysis disappeared in his multivariate analysis where other variables (attainment, gender, centre type, age, exam board and subject) were controlled for.

Centre type

All three studies found the most exactly accurate predictions and the least over-predictions for applicants from independent and selective/grammar schools. Predictions were more accurate and less prone to over-prediction for applicants from state comprehensive schools and sixth-form colleges than for applicants from FE colleges. Also, no centre type was more or less prone to under-prediction. BIS's 2013 report cautioned that the apparent effect of centre type may reflect the association between centre type and attainment.

In Delap's multivariate analysis, compared to comprehensive schools, only FE colleges were found to over-predict significantly more (by 0.28 grade).

Disability

The two BIS reports documented the pattern of accuracy of predictions for applicants who declared different kinds of disability. However, the authors expressed concern about the representativeness of the data, given the low number of applicants declaring any disability in the sample.

Age

All three studies reported that predictions for younger applicants were more accurate and less likely to be over-generous.

However, Delap's multivariate analysis found less difference between predicted and achieved grades (0.18-grade less over-prediction, that is, more accuracy) for 19-year-olds than for 18-year-olds, which contradicts the result of his univariate analysis of age.

Exam board

Delap found some differences in prediction accuracy between exam boards in his univariate analysis, but those differences disappeared in his multivariate analysis.

Subject

UCAS reported for the nine most commonly studied subjects, there was an increase from 2010 to 2015 in the proportion of predicted grades being over-predictions. The highest proportion was consistently found in physics, rising from 55% to 64%, while English literature consistently had one of the lowest proportions of over-predictions, which still rose from 42% to 50%. (The rise in over-predictions might be linked to universities' increasing use of unconditional offers. Teachers' predictions have become apparently less accurate because they have not factored in the influence that unconditional offers from universities can have on some students' study behaviour.)

In Delap's and UCAS's multivariate analyses some subject differences remained. Grades in the nine most popular subjects in Delap's sample were all on average over-predicted; geography grades were the most accurately predicted (with an average 1.13-grade over-prediction) and maths grades the least accurately predicted (with an average 1.73-grade over-prediction).

Studies of grade points

The studies that fall into this category are UCAS (2016), which analysed data from 2010 to 2015 on applicants predicted to achieve at least 12 grade points; and Wyness (2016), which analysed data from 2013 to 2015 on 858,835 applicants. In addition, Shiner and Mahood (2002) studied the effect of ethnicity using data on 7,383 applicants in 1996.

Attainment

UCAS analysed prediction accuracy by GCSE prior attainment. Both their univariate and multivariate analyses showed that the lower the applicant's prior attainment, the more over-predicted the predictions for them turned out to be.

Wyness analysed prediction accuracy by achieved A level attainment and found that the lower the achieved grade points, the greater the over-prediction. (See under *Disadvantage* for an elaboration.)

Gender

Both UCAS and Wyness found little or no difference between predictions for males and females in their univariate analyses. In their multivariate analyses, UCAS found males to have a 1.9% lower probability of having their grade points over-predicted by two or more points, while Wyness found males to have a 0.2% lower probability of having their grade points under-predicted by any amount. The two findings can be seen as converging on more accurate predictions for males.

Ethnicity

Shiner and Mahood found larger differences between predicted and achieved grade points (suggesting less accurate but more optimistic predictions) for Black and Asian applicants than for White applicants. Wyness and UCAS both reported more over-predictions by two or more grade points for Black and Asian applicants than for White applicants. Wyness also noted more exactly accurate predictions and more mild over-predictions (by one grade point) for White applicants.

These effects of ethnicity on prediction accuracy remained in Wyness's and UCAS's multivariate analyses. UCAS found Asian and Black applicants to have a 5.1% and 9.2% higher probability, respectively, of having their grade points over-predicted by two or more grade points, and Wyness found Asian and Black applicants to have a 3.7% and 6.9% lower probability, respectively, of having their grade points under-predicted by any amount.

Centre type

From her univariate analysis, Wyness commented on the high level of accuracy of predictions for applicants from independent schools. The multivariate analysis showed that compared to state schools, independent and grammar schools and academies had a higher probability (averaging about 1.1%), and FE colleges had a 2.4% lower probability, of under-predicting applicants' grade points by any amount.

Disadvantage

Both UCAS and Wyness examined prediction accuracy by POLAR3 classification, an area-based measure of educational disadvantage developed by HEFCE.

UCAS's univariate and multivariate analyses both showed that the more disadvantaged the applicant, the more likely that the prediction for them turned out to be a severe over-prediction (defined as an over-prediction by two or more grade points), but while the univariate analysis suggested the difference in probability of a severe over-prediction between the most and the least disadvantaged applicants was as much as 10%, the multivariable analysis showed the difference to be less than 3%.

Wyness observed that the less disadvantaged the applicant, the more likely that the prediction for them turned out to be exactly accurate or a mild over-/under-prediction by one grade point, and the less likely that the prediction turned out to be a severe over-prediction (that is, over-prediction by two or more grade points). Her multivariate analysis found an interaction between disadvantage and actual attainment. For low-attaining applicants, the more disadvantaged they were, the less prone they were to under-prediction; compared to the least disadvantaged low-attaining applicants, the most disadvantaged low-attaining applicants had a 0.3% lower probability of under-prediction by any amount. For mid-attaining applicants, disadvantage had no effect on their susceptibility to under-prediction. For high-attaining applicants, the more disadvantaged they were, the more prone they were to under-prediction; compared to the least disadvantaged high-attaining applicants, mildly- and highly-disadvantaged high-attaining applicants had a higher probability (averaging about 1.1%) of under-prediction by any amount. In other words, the effect of disadvantage on prediction accuracy was reversed, depending on applicants' level of attainment.

In summary, the one clear finding in this literature is that some levels of attainment are easier to predict accurately than others. There are some effects of centre type and subject on prediction accuracy, but these effects are very small. There are conflicting reports of the effects of gender and age. There is likely a genuine effect of ethnicity (that is, more over-prediction for Black applicants and applicants from some Asian subgroups), but the effect size (expressed as the fraction of a grade) has not been properly estimated. There is also likely a genuine effect of disadvantage or socio-economic status (that is, more over-prediction for the more disadvantaged) or an interaction between disadvantage and attainment (that is, less over-prediction for the more disadvantaged among high attainers), but again the effect size (expressed as the fraction of a grade) has not been properly estimated.

Teacher-estimated grades in awarding

As Dhillon (2002) reports, teacher estimates were previously collected by exam boards for the purposes of standard setting and maintaining. The author notes that ‘research regarding the accuracy of these teachers’ estimates tended to focus on the robust finding that such estimates largely erred on the side of optimism, with teachers over-predicting rather than under-predicting awarded grades’.

However, as Baird (1997) concluded, most of the evidence suggested that centres were good at rank ordering students. For example, Baird found that 75% of correlations between estimates and actual grades *within* centres were above 0.75 in A level biology, English and French, although in sociology 72% were lower than 0.75.

Indeed, research also suggests that accuracy (rather than correlation) varied between qualifications but not in a systematic way. Baird also reports research showing some evidence of bias in favour of female students and older students (17+ at GCSE and 19+ at A level), such that they were estimated higher grades than male students and younger students respectively. In both cases the effects were relatively small (approximately 0.05 of a grade) but statistically significant.

Delap (1994) studied estimated grades in GCSE English, maths, science and geography. The author reported over-prediction of grade Cs and attributed this to teachers being inclined to give students the benefit of the doubt at this crucial borderline. This study also reported a gender effect in English and science (around a tenth of a grade) such that male students were over predicted in science and female students over predicted in English language, as well as a bias favouring older students in these subjects. Additionally, in maths, FE colleges tended to have relatively higher estimates than comprehensive schools. No significant effects of ethnicity or exam board were observed.

Reporting on the accuracy of forecast grades for OCR GCSEs in the summer exam series of 2014, Gill and Benton (2015) observed that while approximately 43% of grades across all subjects were completely accurate, 87% of grades were accurate to within one grade, that is, only 13% of the forecast grades were more than one grade out. The authors also reported that the accuracy of forecast grades differed along the attainment scale such that higher grades were more accurately predicted (for example, 65% accuracy at grade A* and 52% accuracy at grade A compared with 14% and 11% accuracy at grade G and U respectively). As the tendency of predictions was to over-predict (42% of grades) rather than under-predict (14% of grades), the pattern of accuracy across the grading scale may be due to a 'ceiling effect' at the higher end (where, for example, it is not possible to over predict an A* grade). These trends were found to be broadly stable across both academic subjects and school types.

Similarly, Gill and Chang (2013) report on the accuracy of predicted A level grades for the OCR A level summer exam series of 2012, observing that approximately 48% of grades across all subjects were completely accurate and 92% were accurate to within one grade. The accuracy of predicted A level grades also varied across the grading scale such that higher grades were more accurately predicted than lower grades (for example, 63% of predictions were correct at grade A compared with 27% at grade E). This report also examined the accuracy of predictions of A level points score where grades were converted into points as per the UCAS tariff (A*=140, A=120, B=100, C=80, etc.) for students taking three A levels, and found that approximately 28% of scores were predicted accurately and 66% were accurate within 20 points. The authors note that this represents a decline in accuracy compared to the previous year.

For both GCSEs and A levels, it was observed that independent schools and grammar schools made the most accurate and least optimistic predictions (Gill & Benton 2015; Gill & Chang 2013) although the authors suggest that this may be due to their (on average) higher performing students.

In a study of over 7,000 estimated A level grades submitted to a UK exam board, Delap (1995) reported that teachers were reluctant to predict low grades for their students such that only 650 N grades or Us were predicted compared to 1,700 being awarded. This study also applied a multilevel modelling approach to examine the effects of a variety of factors across a range of subjects and found a significant effect of student's gender in three subjects: biology, geography and maths, in each case favouring female students compared with male students achieving the same awarded grade. The author also suggests that a similar gender effect may also have been present in estimates supplied for the other subjects. Age was seen to significantly affect predicted grades in English language and literature and sociology, but the magnitude of this effect was vanishingly small, for example in English language and literature the author notes the effect was equivalent to 'an increase in the average estimated grade about one-tenth of a grade for an increase in age of 10 years'. Conversely, although there was no significant effect of school type in most subjects, the observed (statistically non-significant) effects were relatively large, and the author therefore suggests that 'different school types can apparently play a substantial role in the estimated grade supplied'.

Conclusion

In the first part of this paper, we reviewed some research studies of potential bias in teacher assessment and saw that differences between teacher assessment and exam assessment results can sometimes be linked to student characteristics like gender, special educational needs, ethnicity and age but the effects, when present, are small and inconsistent across subjects.

In the second part, we reviewed research that used data held by UCAS on A levels of multiple exam boards to examine the accuracy of the A level grade predictions made by teachers for their students' university admission applications.

In the third part, we reviewed studies that made use of individual exam boards' data to examine the accuracy of the GCSE and A level grades that exam boards used to collect from teachers.

The latter two strands of work are of greater relevance to our present purpose. Although lower grades may be under-represented in the UCAS data compared to individual boards' data (because those with lower grades are less likely to apply to universities and hence do not feature in UCAS data), the two strands of work found the same distribution of exactly accurate and over-/under-predictions and pattern of grade-/attainment-dependent prediction accuracy.

The findings about individual variables are broadly similar: subject has a small but unsystematic effect; gender and age have small effects which are inconsistent across subjects; centre type has a small effect which can be speculated to be attributable to the correlation between centre type and attainment and attainment-dependent prediction accuracy. There are likely some effects on prediction accuracy of ethnicity (that is, more over-prediction for some ethnic minority groups) and disadvantage (that is, more over-prediction for the more disadvantaged in general, and less over-prediction for the more disadvantaged among high attainers) but those effects have not been properly estimated.

References

Baird, J. A. (1997). *Teachers' Estimates of A level Performance*. AEB Internal Report.

Baird, J. A. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, *40*(2), 191–202. https://doi.org/10.1080/0013188980400207

BIS. (2011). *Investigating the accuracy of predicted A level grades as part of the 2009 UCAS admission process*. BIS. 2011-Investigating-accuracy-predicted-a-level-grades.pdf

BIS. (2013). *Investigating the accuracy of predicted A level grades as part of the 2010 UCAS admission process*. BIS. 2013-Investigating-the-accuracy-of-predicted-A-level-grades.pdf

Delap, M. R. (1994a). An investigation into the accuracy of A-level predicted grades. *Educational Research*, *36*, 135–148.

https://doi.org/10.1080/0013188940360203

Delap, M. R. (1994b). Multilevel analysis of teachers' estimates of candidates' GCSE results - Results of the 1993 pilot study. *RAC/623*.

Delap, M. R. (1995). Teachers' Estimates of Candidates' Performances in Public Examinations. *Assessment in Education: Principles, Policy & Practice*, *2*(1), 75–92. https://doi.org/10.1080/0969594950020106

Dhillon, D. (2002). AS-LEVEL PSYCHOLOGY B — SUMMER 2001: Did teachers underestimate candidates' grades?

Dhillon, D. (2005). Teachers' estimates of candidates' grades: Curriculum 2000 advanced level qualifications. *British Educational Research Journal*, *31*(1), 69–88. https://doi.org/10.1080/0141192052000310038

Fajardo, D. M. (1985). Author race, essay quality and reverse discrimination. *Journal of Applied Social Psychology*, *15*, 255–268. https://doi.org/10.1111/j.1559-1816.1985.tb00900.x

Gibbons, S., & Chevalier, A. (2008). Assessment and age 16 + education participation. *Research Papers in Education*, *23*(2), 113–123. https://doi.org/10.1080/02671520802048638

Gill, T., & Benton, T. (2015). The accuracy of forecast grades for OCR GCSEs in June 2014: Statistics Report Series No.91. *Statistics Report Series No.91.* Cambridge Assessment. *https://www.cambridgeassessment.org.uk/Images/374825-statistical-reports.pdf*

Gill, T., & Chang, Y. (2013). The accuracy of forecast grades for OCR A levels in June 2012. *Statistics Report Series No.64.* Cambridge Assessment. *https://www.cambridgeassessment.org.uk/Images/374825-statistical-reports.pdf*

Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology*, *76*(5), 777–781. https://psycnet.apa.org/doi/10.1037/0022-0663.76.5.777

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgements of academic achievement: a review of literature. *Review of Educational Research*, *59*(3), 297–313. https://doi.org/10.3102%2F00346543059003297

Krkovic, K., Grei, S., Kupiainen, S., & Vainikainen, M. (2014). Teacher evaluation of student ability: what roles do teacher gender, student gender, and their interaction play? *Educational Research*, *56*(2), 244–257. https://doi.org/10.1080/00131881.2014.898909

Marcenaro-gutierrez, O., & Vignoles, A. (2014). A comparison of teacher and test-based assessment for Spanish primary and secondary students. *Educational Research*, *57*(1), 1–21. https://doi.org/10.1080/00131881.2014.983720

Mastergeorge, A. M., & Martínez, J. F. (2010). Rating Performance Assessments of Students With Disabilities: A Study of Reliability and Bias. *Journal of Psychoeducational Assessment*, *28*(6), 536–550.

https://doi.org/0734282909351022

Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The Relationship between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2 , 1996 – 98. *British Educational Research Journal*, *27*(2). https://doi.org/10.1080/0141192012003710

Scottish Examination Board. (1992). *Investigation into the effects of the characteristics of candidates and presenting centres on possible marker bias*. Scottish Examination Board.

Shiner, M., & Modood, T. (2002). Help or hindrance? Higher education and the route to ethnic equality. *British Journal of Sociology of Education.*, *23*, 209–232. https://doi.org/10.1080/01425690220137729

Spear, M. G. (1984). Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education*, *6*(4), 368–377. https://doi.org/10.1080/0140528840060407

Thomas, S., Smees, R., Madaus, G. F., Raczek, A. E., Thomas, S., Smees, R., … Raczek, A. E. (1998). Comparing Teacher Assessment and Standard Task Results in England: the relationship between pupil characteristics and attainment. *Assessment in Education: Principles, Policy & Practice*, *5*(2), 213–246. https://doi.org/10.1080/0969594980050205

UCAS. (2016). *Factors associated with predicted and achieved A level attainment*. UCAS. https://www.ucas.com/file/71796

Wyness, G. (2016). *Predicted grades: accuracy and impact*. UCU. https://www.ucu.org.uk/article/8558/Predicted-grades-accuracy-and-impact

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

April 2020 **Ofqual/20/6611**