RESEARCH AND ANALYSIS

Systematic divergence between teacher and test-based assessment: literature review

ofqual

Contents

Authors

This report was written by Ming Wei Lee and Paul Newton, from Ofqual's Strategy Risk and Research directorate.

Executive Summary

Ofqual's equalities analyses last year concluded that the centre assessment grades (CAGs) of summer 2020 did not systemically disadvantage students on the basis of their protected characteristics or socioeconomic status, suggesting that the teacher judgements/predictions underlying CAGs in 2020 did not differ from the mainly exam-based assessments in 2018 and 2019 in susceptibility to bias. Given the conceptual distinction between teacher prediction of prospective performance and teacher assessment of current attainment, a further literature review was conducted this year on systematic divergence between results from teacher and test-based assessments, to raise awareness of potential risks to the dependability of assessment results which are based entirely on teacher judgements.

The reviewed literature suggests that the relative agreement (as measured by the similarity in the rank order of students) between results from teacher and test-based assessments is of a comparable level to the relative agreement between teacher prediction and actual achievement. While there is ample evidence of teachers' tendency towards generosity in grade prediction, there seems little evidence of the equivalent (that is, teacher over-rating relative to test results) in teacher assessment in National Curriculum assessment in England, but studies from abroad found that in case of disagreement between teacher and test grades, over-rating by teachers (relative to the test grades) was much more common than under-rating.

With respect to teacher assessment, evidence of teacher bias in relation to gender is mixed, but a slight bias in favour of girls (or against boys) is a common finding. Evidence in relation to ethnicity is also mixed: there are findings of bias against as well as in favour of each minority group (relative to the majority group) and findings of no bias. Evidence on disadvantage and special educational needs (SEN) is less mixed, with bias against the more disadvantaged (or in favour of the less disadvantaged) and bias against pupils with SEN (or in favour of those without) being common findings.

The unique circumstance under which teacher judgements are called upon for summer 2021 means that the teacher assessment to be conducted has no exact parallel in the 22 studies reviewed, and no study reviewed is unreservedly informative about summer 2021. By conducting this review, we are not suggesting that only teacher assessment is open to bias or that evidence of systematic divergence is incontrovertible proof of error or bias in teacher judgements. The greater subjectivity of teacher assessment, however, makes it more vulnerable to bias than test-based assessment, and repeated reports of evidence of systematic divergence therefore highlight the possibility of bias in teacher assessment judgements. It will therefore be important for teachers to be aware of the potential

risks to the validity of their judgements, and take steps to mitigate them by following Ofqual's objectivity guidance.

Introduction

In April 2020, as part of its equality impact assessment of the then proposed method of summer awarding based on standardisation of centre assessment grades (CAGs), Ofqual conducted a rapid review of the research literature to understand the nature and extent of any bias that might arise in using those grades (Lee & Walter, 2020). The review covered research on:

(i) teacher assessment in general

(ii) teacher-predicted A level grades used in university admission

(iii) teacher-forecast grades that exam boards used to use as a source of evidence to support the setting of grade boundaries

(ii) and (iii) were of greater relevance at the time, because it was indeed grade prediction and forecasting that teachers were to engage in – for summer 2020, CAGs were the grades that centres predicted or forecast their students to have achieved if they had sat the exams.

One year on, summer exams have again been cancelled because of the pandemic, and following government policy, students will be issued grades based on teachers' judgements. For summer 2021, teachers will not be asked to predict or forecast grades. Instead, they are expected to make summative judgements about their students' attainment in relation to the subject content that has been taught, drawing on a range of assessment evidence as they see fit. Given the difference in the task required of teachers, it is opportune to revisit last year's literature review and update the part on teacher assessment, which takes on greater relevance in the present context.

By conducting reviews of the research literature on possible biases in teacher judgements, we are not suggesting that only teacher assessment is susceptible to bias. In fact, our equalities analyses last year (Lee, Stringer & Zanini, 2020) concluded that the CAGs of summer 2020 did not systemically disadvantage students on the basis of their protected characteristics or socioeconomic status (and neither did the standardised/calculated grades nor the final grades, which were the higher of CAGs and calculated grades), suggesting that the teacher judgements/predictions underlying CAGs in 2020 did not differ from the mainly exam-based assessments in 2018 and 2019 in susceptibility to bias. Test-based as well as teacher assessment results can show achievement gaps by student characteristics. By themselves they do not indicate bias in the assessment, as they may reflect genuine differences in attainment. Our equalities analyses last year found in the GCSE grades of summer 2019 (when the grades in all subjects were wholly or mainly based on exams) educationally significant achievement gaps along the lines of gender, socioeconomic status, and so on that could not be explained by gaps in prior attainment. Because it is hard for biases to rear their heads in the relatively more objective administration, marking and grading of exams, those achievement gaps are much more likely to reflect genuine differences in attainment resulting from societal inequalities in educational opportunities than biases. Having said that, we do keep an open mind about bias in test-based assessment. And we are aware of concerns about the reliability of results of qualifications (which have

only or mainly test-based assessments) (Bramley & Dhawan, 2010; Rhead, Black & Pinot de Moira, 2016, 2018; Wheadon & Stockford, 2010).

There are a number of widely cited literature reviews on teacher assessment. Brookhart (2013) summarised research in the American context. Hoge and Coladarci (1989; see also re-analysis by Kaufmann, 2020), Südkamp, Kaiser and Möller (2011), and Urhahne and Wijnia (2021) carried out meta-analyses of studies of the accuracy of teacher judgement of academic achievement and recounted findings of moderating effects (or the lack thereof) of student characteristics on teacher judgement accuracy. They included almost no British study. Reviews that considered relevant British studies were provided by Harlen (2005) and Johnson (2013) who, however, did not distinguish between teacher judgement of current attainment and teacher estimation of prospective performance. Because of the limitations of the extant literature reviews, we conduct our own review of the primary literature on systematic divergence between results from teacher assessments and test-based assessments. Evidence of systematic divergence does not provide direct evidence of error or bias in teacher assessments. This is because it is always possible, for example, that the comparator test might be biased (as suggested above), or that both are biased.[1] However, the greater subjectivity of teacher assessment makes it more vulnerable to bias than test-based assessment, and evidence of systematic divergence therefore points to the possibility of bias in teacher assessment judgements. By highlighting any such evidence, we are not suggesting that the teacher judgements for summer 2021 will necessarily be biased, but we hope to raise awareness of potential risks to the dependability of those judgements and the need to mitigate those risks.

We identified 22 relevant studies from the past 30 years for this review. They satisfied all our inclusion criteria:

(i) the study contains some analysis of the results of teachers' summative assessment of their students' attainment in a subject at a particular time

(ii) the study contains a comparison of contemporaneous teacher and test-based assessment results on the same students

(iii) the comparison in (ii) is quantitative in nature, employs suitable statistical methods and draws on large-scale naturalistic or semi-naturalistic data[2]

The studies can be grouped into 3 types based on their data source:

(i) research using naturalistic data from statutory National Curriculum assessment in England

(ii) research similar to (i) from other jurisdictions

(iii) research using semi-naturalistic data collected specially for large-scale cohort studies or research projects in the UK or abroad

[1] We shall discuss other caveats in interpreting evidence of (systematic) divergence (or the lack thereof) throughout this paper.

[2] It is worth pointing out what studies are not included in this review because of the inclusion criteria. We have not included work on formative teacher assessment, teacher prediction, other judgements or decisions that teachers and schools make on pupils, such as set allocation (see, for example, Connolly, Taylor, Francis, Archer, Hodgen, Mazenod & Tereshchenk, 2019), tier entry (see, for example, Strand, 2012), school exclusion (see, for example, Department for Education, 2019), differences between coursework and exam assessments (see, for example, Pinot de Moira, 2020), and experimental studies of teacher assessment (see, for example, Malouf & Thorsteinsson, 2016 for a meta-analysis of experimental studies of non-blind vs. blind marking).

Evidence from National Curriculum assessment in England

Statutory National Curriculum assessments (NCA) in England include a combination of teacher-led and test-based assessments completed at the end of each Key Stage. Studies that investigated teacher assessment using NCA data are:

- Burgess and Greaves (2013): KS2 (year 6, age 11) in 2002-5
- Durant (2003): KS1 (year 2, age 7) in 1998-2002; KS2 (year 6, age 11) in 1996-2002; KS3 (year 9, age 14) in 1998-2002
- Gibbons and Chevalier (2008): KS3 (year 9, age 14) in 2002-5
- Plewis (1997): KS1 (year 2, age 7) in 1991
- Reeves, Boyle and Christie (2001): KS2 (year 6, age 11) in 1996-8
- Rimfeld, Malanchini, Hannigan, Dale, Allen, Hart and Plomin (2019): KS1/2/3 (year 2/6/9, age 7/11/14) in 2001-3/2005-7/2008-10
- Thomas, Smees, Madaus and Raczek (1998): KS1 (year 2, age 7) in 1992

At the time the data that these studies drew on was collected, teacher assessment and test results carried the same status. The tests captured a snapshot of pupils' attainment and the teacher assessments took account of evidence of attainment gained over the programme of study. The 2 assessments were supposed to provide complementary information about pupils' attainment at the end of a Key Stage, although some curricular elements (for example, speaking and listening in English) were, by design, only assessed by teachers. When interpreting any finding of (systematic) divergence between results from the 2 assessment methods, one should bear in mind the possibilities that the 2 assessment methods did not measure the exact same construct and that pupils could be genuinely different at the times of the assessments (for example, they could have been motivated and enthused not to the same degree by the 2 assessment methods, or they could have improved their attainment in the intervening time between the 2 assessments).

NCA as a data source for researching teacher and test-based assessments has other shortcomings, which were discussed in most detail by Reeves et al. (2001) and mentioned in only some of the studies cited above. The structure of the NCA system was such that the teacher and test-based assessments were not entirely independent. The teacher responsible for the teacher assessment could limit, to some extent, what their student could achieve on the test, by determining, for example, the tier to enter the student for in KS3 maths and science, whether the student should enter the extension tests for the higher levels at KS2, that the student was performing at too low a level to take the test. What's more, the teacher had access to their students' test results when finalising their judgements of the students.[3] NCA results from the 2 assessment methods could have a higher level of agreement than would have been possible if they had operated more independently,

[3] This is not true of the very early KS1 data he analysed, according to Plewis (1997).

which should be borne in mind when interpreting any finding of lack of (systematic) divergence.

In NCA, teacher and test-based assessments were reported on the same scale of attainment levels, so it was possible to examine the absolute agreement between results from the 2 assessment methods, that is, how often a pupil was judged to be performing at the same level by the teacher as by the test (see Table 1 for a summary of the relevant analyses).[4] The rows labelled TA=Test in Table 1 show the levels of absolute agreement in different subjects at different Key Stages. Over 60% absolute agreement was found in all subjects at KS2 and KS3, except for KS3 English, while all curriculum areas at KS1 and KS3 English saw levels of absolute agreement of under 60%. When the levels from the teacher and the test did not match, teacher under-rating relative to test results was slightly more common than over-rating in all analyses of science, but there was no clear tendency towards teacher under- or over-rating in English and maths.[5]

Table 1. *Summary of analyses of absolute agreement between results from teacher and test-based assessments in National Curriculum assessment*

	Teacher-assessed level (TA) = Test-assessed level (Test)?	KS1* (Level 0-4, with 3 sublevels of Level 2)	KS2** (Level <2 to 5/6)	KS3* (Level 0-8)
English	TA = Test	Reading: 58% Writing: 49%	71-76%	53%
	TA < Test	Reading: 26% Writing: 20%	14-15%	22%
	TA > Test	Reading: 16% Writing: 32%	8-11%	24%
Maths	TA = Test	50%	76-79%	69%
	TA < Test	26%	9-10%	18%
	TA > Test	23%	10-13%	13%
Science	TA = Test		72-74%	63%
	TA < Test		15-20%	21%
	TA > Test		7-12%	16%

* Based on Durant (2003, Annexes 1-3 for KS1 and 7-9 for KS3).

[4] In research on predicted/forecast grades, teacher prediction/forecast was always on the same scale as exam grades or grade-points, which made it easy to assess absolute agreement (or absolute accuracy, relative to actual achievement) and any tendency towards over- or under-prediction. In research on teacher assessment, teacher and test-based assessments were not always on the same scale, and analyses of absolute agreement are hard to come by. We can find only two analyses of absolute agreement from abroad, apart from the ones summarised in Table 1. The footnotes to Table 1 give the sources for the summary. Note that Plewis (1997), Thomas et al. (1998) and a parliamentary answer from 2009 https://hansard.parliament.uk/Commons/2009-02-26/debates/6a7d6e6b-4b91-4339-8943-28cc33e5271a/WrittenAnswers [under National Curriculum Tests] also contained analyses of absolute agreement at KS1 and KS3, but the way the analyses were organised and presented makes it hard to combine their results with Durant's (2003) in the summary.

[5] Urhahne and Wijnia (2021) reported that their meta-analysis of mostly European studies "strongly supported the hypothesis that teachers overestimate student achievement on a standardized test" (p.6). Their talk of such overestimation being "motivationally favorable" suggests that the studies they meta-analysed were on teacher predictions or formative teacher assessment rather than summative teacher assessment.

** Based on Burgess and Greaves (2013, Table 1), Durant (2003, Annexes 4-6) and Reeves et al. (2001, Table VI).

Rimfeld et al. (2019) and Thomas et al. (1998) calculated simple correlations between teacher assessment and test results. These correlations can be taken as measures of relative agreement between the 2 assessments, that is, how well teachers' rank ordering of pupils matched the test's rank ordering of the same pupils. According to Rimfeld et al., the correlations were 0.74 for both English and maths at KS1 and ranged from 0.64 to 0.78 for all areas at KS2 and KS3, except for KS3 science. The low correlation of 0.25 for KS3 science was likely due to the low reliability of the test, according to Rimfeld et al.

The moderate level of absolute agreement shown in Table 1 suggests there was non-negligible divergence between results from teacher and test-based assessments in NCA. The pattern of divergence varied by attainment level, as noted by Burgess and Greaves (2013) and Gibbons and Chevalier (2008): in all subjects, higher attainers were more susceptible to under-rating (relative to test results) by teachers than lower attainers, and lower attainers were more likely to be over-rated by teachers than higher attainers. Note that because of the floor/ceiling of the grade scale, teacher ratings for the lowest/highest possible attainers could only be the same as, or be over-/under-ratings relative to test results. The variation of the pattern of divergence by attainment level was not an artefact of the floor/ceiling of the grade scale and did not pertain just to the highest and lowest possible attainers, according to Gibbons and Chevalier's analysis.

The issue of any other systematic pattern of divergence in NCA data was investigated in 5 of the 7 studies listed above. The 5 studies did not employ the same method of analysis to test for systematic divergence. Their common rationale was that the lesser subjectivity of the test makes it less vulnerable to bias and therefore test results are likely to better reflect students' attainment, and if any achievement gap by a student characteristic is found in teacher-assessed results after controlling for (or conditioning on) the achievement gap by the same student characteristic in test results, it constitutes evidence of systematic divergence which points to possible bias in teacher assessment judgements in relation to that student characteristic.[6]

The student characteristics this review focuses on are gender, ethnicity, socioeconomic status, special educational needs (SEN) status and English as an additional language (EAL) status. Tables 2 to 4 show the conditional achievement gaps in teacher-assessed results that were found or not found in the various analyses, based on significance testing of the relevant statistics (p<.05) from the sources given on the bottom row of the tables. Also given on the bottom row is the number of observations in the analysis that the cited significant test results came from. One should be mindful that statistical significance is partly dependent on sample size: a genuine effect may fail to be found significant because of insufficient sample size, while an effect which is too small to be practically meaningful may be found significant in an analysis of a very large sample.

[6] Gibbons and Chevalier (2008) did not assume test results better reflect attainment. They analysed the difference between KS3 teacher and test levels, controlling for the average of KS3 teacher and test levels or for KS2 prior attainment.

All the achievement gaps shown in Tables 2-4 were conditional on test results, and, with the exception of those from Plewis (1997), they were conditional additionally on other student characteristics, and in some cases, also on schools. The notion of conditioning can be illustrated with an example using real data. In Burgess and Greaves's (2013) analysis of KS2 English data, one can see a 'raw' difference in teacher-assessed results between Indian and white pupils in that Indian pupils had a 2.1 percentage point higher probability than white pupils of being under-rated by teachers relative to their test level. Given the finding mentioned above that high attainers were more likely to be under-rated by teachers than low attainers, the raw difference between Indian and white pupils could in part be due to higher attainment on average of Indian pupils. After conditioning on test results (which arguably better reflect attainment), the difference in probability of teacher under-rating between Indian and white students reduced slightly to 1.7 percentage point, which points to a statistically significant conditional achievement gap in teacher-assessed results in favour of white over Indian pupils. In a further analysis, the difference reduced to virtually zero after conditioning on other student characteristics including SEN status, EAL status, free school meal eligibility, gender and 'tested in wrong year' status. The further analysis tells us that there was no achievement gap between Indian and white pupils in teacher-assessed results, considering the systematic differences between the groups in those other student characteristics, the correlations between those characteristics and attainment, and the correlation between attainment and the probability of teacher under-rating. A still further analysis factored in, in addition, the fact that schools differed in the tendency to under-rate pupils relative to their test results. After conditioning on schools, the difference in probability of teacher under-rating between Indian and white students became -1.8 percentage point, which points to a significant within-school conditional achievement gap in teacher-assessed results, now in favour of Indian over white students.[7] All in all, the raw achievement gap by a student characteristic is often the combined influence of that characteristic and other factors. The statistical technique of conditioning helps bring to light as pure as possible an achievement gap in relation to a student characteristic which is net of the effects of any correlating characteristics.

In Tables 2-4, the notation for a conditional achievement gap is A>B, which indicates that for members of A and B who achieved the same level at the test, teachers rated members of A more highly than members of B on average. It cannot be inferred, however, whether, relative to test results, A was over-rated while B was under-rated or 'correctly' rated, or A was 'correctly' rated while B was under-rated, or A and B were both under-rated or both over-rated, but to statistically significantly different degrees.

[7] The result of the additional 'conditioning on schools' analysis suggests that some schools under-rated more of their pupils in general than others and that Indian pupils were more likely to be in schools that had a greater tendency to under-rate pupils in general. Note that the rather drastic effect of the additional conditioning on schools seen with the Indian data was not observed very often. In Burgess and Greaves's analyses, the additional conditioning on schools had minimal effect for most conditional attainment gaps, meaning those gaps happened within schools and were not driven by differences between schools.

Table 2. *Conditional achievement gaps by student characteristics in teacher assessment in National Curriculum assessment in English*

	Burgess & Greaves (2013) [KS2]	Gibbons & Chevalier (2008) [KS3]	Plewis (1997) [KS1]	Reeves et al. (2001) [KS2]	Thomas et al. (1998) [KS1]
Gender	Female > Male	Male > Female	Female > Male	Mostly no significant gap, but sometimes Male > Female	Female > Male
Ethnicity	White > All black groups, Pakistani, Other Asian, Mixed white and black Caribbean, Other Indian, Chinese, Mixed white and Asian > White No significant gap between White and Bangladeshi, Mixed white and black African, Mixed other	White > Asian No significant gap between White and Black, Mixed, Other	White > Non-white (combining African and African Caribbean, Indian, Pakistani)		
Socio-economic status	NoFSM > FSM	No significant gap	Higher social class > Lower social class [social class based on school postcode]		NoFSM > FSM
SEN	NoSEN > SEN			Mostly no significant gap, but sometimes NoSEN > SEN	NoSEN > SEN
EAL	NotEAL > EAL	No significant gap		Mostly no significant gap, but sometimes NotEAL > EAL	NotEAL > EAL
Based on significance testing of relevant statistics in: (number of observations in analysis in brackets)	Specification 4 of online-only Table A3 (2255382)	Column 3 of Table 4 of working paper (1439409)	Table 6 (<7400)	Table VII (between 1203 and 2298)	Model D of Table A3.1 (16840)

Table 3. *Conditional achievement gaps by student characteristics in teacher assessment in National Curriculum assessment in maths*

	Burgess & Greaves (2013) [KS2]	Gibbons & Chevalier (2008) [KS3]	Plewis (1997) [KS1]	Reeves et al. (2001) [KS2]	Thomas et al. (1998) [KS1]
Gender	[not presented]	Female > Male	Female > Male	Female > Male	No significant gap
Ethnicity	White > Black Caribbean, Black African Chinese, Indian, Other Asian, Mixed white and black African, Mixed white and Asian, Mixed other > White No significant gap between White and Black other, Pakistani, Bangladeshi, Mixed white and black Caribbean, Other	Asian, Black > White No significant gap between White and Mixed, Other	White > Non-white (combining African and African Caribbean, Indian, Pakistani)		
Socio-economic status	[not presented]	No significant gap	Higher social class > lower social class [social class based on school postcode]		NoFSM > FSM
SEN	[not presented]			NoSEN > SEN	NoSEN > SEN
EAL	[not presented]	No significant gap		No significant gap	No significant gap
Based on significance testing of relevant statistics in: (number of observations in analysis in brackets)	Specification 4 of Table 4 (2255382)	Column 9 of Table 4 of working paper (1439409)	Table 6 (<7400)	Table VII (between 1206 and 2313)	Model D of Table A3.3 (16840)

Table 4. *Conditional achievement gaps by student characteristics in teacher assessment in National Curriculum assessment in science*

	Burgess & Greaves (2013) [KS2]	Gibbons & Chevalier (2008) [KS3]	Plewis (1997) [KS1]	Reeves et al. (2001) [KS2]	Thomas et al. (1998) [KS1]
Gender	[not presented]	Female > Male	Female > Male	Mostly no significant gap, but sometimes Female > Male	No significant gap
Ethnicity	White > Black African, Black Caribbean, Black other, Mixed white and black Caribbean, Other Indian, Bangladeshi, Chinese, Mixed white and Asian, Mixed other > White No significant gap between White and Pakistani, other Asian, Mixed white and black African	Asian, Black > White No significant gap between White and Mixed, Other	White > Non-white (combining African and African Caribbean, Indian, Pakistani)		
Socio-economic status	[not presented]	No significant gap	Higher social class > lower social class [social class based on school postcode]		NoFSM > FSM
SEN	[not presented]			NoSEN > SEN	NoSEN > SEN
EAL	[not presented]	No significant gap		No significant gap	No significant gap
Based on significance testing of relevant statistics in: (number of observations in analysis in brackets)	Specification 4 of Table 5 (2255382)	Column 6 of Table 4 of working paper (1439409)	Table 6 (<7400)	Table VII (between 1220 and 2307)	Model D of Table A3.1 (9371)

What can we conclude from Tables 2-4? On gender, divergence in favour of girls or against boys was the more common finding, but divergence in the opposite direction was sometimes observed in English. On ethnicity, it is hard to make a systematic comparison across studies because of the different ethnicity groupings used in the studies. It appears for most ethnic groups (compared to white), divergence in one direction was a common finding, but there was at least one occasion where it was either not found or divergence in the opposite direction was found. On socio-economic status, divergence in favour of the less disadvantaged or against the more disadvantaged was commonly, but not always, found in all subjects. On SEN, divergence against SEN pupils or in favour of those without SEN was always found in maths and science and sometimes found in English. On EAL, divergence against EAL pupils or in favour of non-EAL pupils was sometimes found in English, but not in maths or science.[8]

Evidence from other jurisdictions

Comparisons of teacher and test-based assessments in other jurisdictions can be found in the following papers:

- Falch and Naper (2013): data provided by Statistics Norway (maths, English and Norwegian; 10th graders [end of compulsory schooling] in 2002-5)
- Feron, Schils and ter Weel (2016): data from almost all schools in the Limburg region of the Netherlands (combined performance in maths, reading, study skills and science; 6th graders [end of primary education] in 2009)
- Lavy (2008): data provided by Israeli Ministry of Education (multiple science and humanities subjects; 10th to 12th graders in Jewish secular schools taking matriculation exams in 2000-2)
- Lindahl (2016): data provided by Swedish Agency for Education (maths; 9th graders [end of compulsory schooling] in 2002-5)
- Marcenaro-Gutiérrez and Vignoles (2015): Andalusian Social Survey with linked data from Andalusian Educational Authority and regional educational authorities (reading and maths; 11 and 15-year-olds [end of primary and secondary school respectively] in 2010)
- Rangvid (2015): data from administrative registers hosted by Statistics Denmark (multiple science and humanities subjects; 9th graders [end of compulsory schooling] in 2005-11)

All these studies analysed large-scale naturalistic data like England's NCA data. In all but the Spanish study, the assessments that the data pertained to likely held higher stakes for the students than NCA did in England. What teacher assessment entailed was not the same among these studies. For example, in the Israeli system, teacher assessment was the school exam, which differed from the test-based assessment in being internally set and non-blindly marked by teachers. In the

[8] Variation of educational attainment by SEN type (see, for example, Department for Children, Schools and Families, 2009) and by proficiency in English (see, for example, Strand & Hessel, 2018) suggests there is much heterogeneity among SEN pupils and among EAL pupils. We cannot tell from the literature reviewed whether the findings about SEN and EAL status held equally for all SEN subgroups and for EAL pupils with varying levels of proficiency in English.

Norwegian system, teacher grading was supposed to give the highest weight to performance at a final school test, structured identically to, and conducted a few weeks before, the central exit exam (which constituted the test-based assessment in the study), while also taking into account performance throughout the whole school year. In the other systems, teacher assessment was based on the teacher's experience and interaction with the pupil and could draw on all available information. The test-based assessment's vulnerability to bias, relative to teacher assessment, also differed among the studies. The state exams in the Israeli system were likely the least vulnerable in being externally set and externally marked. In contrast, in the Swedish system, the national tests were marked by schools and teachers were allowed (though not encouraged) to mark their own students' answers, and in the Danish system, exam marking was partly done by pupils' own teachers and was non-blind in that pupils' and schools' names were visible to (external) markers. As in England's NCA system, test results were available to teachers at the time teacher assessment judgements were finalised in the Swedish and Dutch systems, but not in the other systems. It should be borne in mind that these system-level differences may increase or decrease the likelihood of finding (systematic) divergence between results from teacher and test-based assessments in individual systems.

Two of the studies provided rare analyses of absolute agreement. In the Dutch case where teachers were aware of test results when finalising teacher assessment judgements, the level of absolute agreement was 82% for males and 80% for females, on an 8-level scale (Feron et al.'s 2015 working paper, Table C1). In the Norwegian case where teacher judgements were made without knowledge of central exit exam results, it was 61% for males and 59% for females, on a 6-level scale (Falch & Naper, 2013: Tables 2 and 3). Falch and Naper's cross-tabulations showed also some dependency of absolute agreement on attainment level: the level of absolute agreement was about 70% for high attainers who scored one of the two highest levels at exams, but about 55% for low attainers who scored one of the two lowest levels at exams. In both the Dutch and Norwegian data, when teacher and test results did not agree, teacher results were much more likely to be over-ratings than under-ratings relative to test results, but note that the Norwegian analysis found that the over-rating tendency did not apply to high, but not the highest, attainers: for high attainers scoring the second highest level at exams, the level of absolute agreement was high at about 71%, but in case of disagreement, teacher ratings were more likely to be under-ratings than over-ratings relative to exam results. This echoes the finding in NCA studies that higher attainers were more susceptible to under-rating by teachers relative to test results.

All studies employed analysis methods that allowed conditional achievement gaps by student characteristics to be examined. Table 5 shows the conditional achievement gaps in teacher assessment that were found or not found in the various studies. Although there is no strong basis on which to make international comparisons, some similarities and differences to the findings in England's NCA can be noted. In relation to gender, divergence in favour of girls or against boys was even more commonly found than in England. On ethnicity or (im)migrant status, the findings were mixed in that there were findings of divergence against as well as in favour of minority groups and findings of no systematic divergence. On socio-economic status, there were findings of divergence in favour of the more disadvantaged or against the less disadvantaged, which was seldom observed in analyses of England data.

Table 5. *Conditional achievement gaps by student characteristics in teacher assessment in other jurisdictions*

	Falch & Naper (2013) [Norway; 10th graders; maths, English and Norwegian combined]	Feron et al. (2016) [The Netherlands; 6th graders; combined performance in multiple areas]	Lavy (2008) [Israel; 10th -12th graders; only maths and English here]	Lindahl (2016) [Sweden; 9th graders; maths]	Marcenaro-Gutiérrez & Vignoles (2015) [Spain; only 15-year-olds' analysis here; reading and maths]	Rangvid (2015) [Denmark; 9th graders; multiple subjects combined]
Gender	Female > Male	Female > Male	Female > Male	Female > Male	No significant gap in reading Female > Male in maths	Female > Male
Migrant/Immigrant status or Ethnic origin	Second generation immigrant > Native No significant gap between First generation immigrant and Native	No significant gap between Native and those who were born, or whose mother or father was born, outside Limburg or abroad	No significant gap between Israeli origin and Non-Israeli origin Recent immigrant > Non-immigrant	No significant gap between Nordic-born and Non-Nordic-born	No significant gap between Non-immigrant and Immigrant	Non-migrant > Migrant
Socio-economic status	Higher SES > lower SES (with parental education level and mother's income level as SES indicators)	More disadvantaged > Less disadvantaged (with father's ability to work as indicator) Less disadvantaged > More disadvantaged (with mother's ability to work and employment status as indicators) No significant gap between the More and Less disadvantaged (with parents' education levels and working pattern as indicators)	Lower SES > Higher SES in English and no significant gap between Lower SES and Higher SES in maths (with father's schooling as SES indicator) No significant gap between Lower SES and Higher SES (with mother's schooling as SES indicator)		Lower SES > Higher SES (with school type and higher cultural index as SES indicators) No significant gap between Higher and Lower SES (with parental education level and average cultural index as SES indicators)	Higher SES > Lower SES
Based on significance testing of relevant statistics in: (number	All subjects column under Models in	Columns 1-3 in Table C2 of working paper (1100)	Columns 3 and 7 of Table 8 (assumed to be up to 109928 [maths], up to 84850	Column 3 of Table 3 (268325)	Table 2 (between 1041 and 1114 [reading], between	Specification 4 of Table 2 (4233824)

of observations in analysis in brackets)	Table 9 in Table A2 (130464)		[English]; number of observations was twice the number of students)		1011 and 1081 [writing])	

Evidence from large-scale cohort studies or research projects in the UK or abroad

The studies reviewed so far analysed data from real assessments that took place in the respective education system. There are other relevant studies where either the teacher or the test-based assessment results or both came about less naturalistically:

- Campbell (2015): fourth sweep of UK Millennium Cohort Study (reading and maths; when members were 7 years old in 2007)
- Cornwell, Mustard and Van Parys (2013): Early Childhood Longitudinal Survey – Kindergarten Cohort in the US (reading, maths, science; 1st/3rd/5th graders in 2000/2002/2004)
- Hansen (2016): second sweep of UK National Child Development Study (general ability; when members were 11 years old in 1968/9)
- Johansson, Myrberg and Rosén (2012): Swedish PIRLS (Progress in International Reading Literacy Study) and its national extension (reading; 3rd and 4th graders in 2001)
- Martínez, Stecher and Borko (2009): Early Childhood Longitudinal Survey – Kindergarten Cohort in the US (maths; 3rd/5th graders in 2002/2004)
- Meissel, Meyer, Yao and Rubie-Davies (2017): Consortium for Professional Learning project in New Zealand (reading and writing; 8- to 13-year-olds in 2012 and 2013)
- Perkins, Kleiner, Roey and Brown (2004): The High School Transcript Study (HSTS) linked with The National Assessment of Educational Progress (NAEP) in the US (maths and science; 12th graders graduating in 2000)
- Ready and Wright (2011): Early Childhood Longitudinal Survey – Kindergarten Cohort in the US (literacy; kindergarteners in 1999)
- Shackleton and Campbell (2014): fourth sweep of UK Millennium Cohort Study (reading and maths; when members were 7 years old in 2007)

With the exception of the New Zealand study and the linked HSTS/NAEP study, the teacher assessments analysed in these studies did not appear to have any official status and were produced for the purpose of the relevant cohort study or research project. In all studies, there is no reason to doubt the quality of the test instruments, but the test-based assessments were conducted for a research purpose and were very low-stakes for the pupils. While there may be question marks over the authenticity of the assessment data in these studies, it has been argued that teacher assessments made outside the education and assessment system better reflect what teachers think about their pupils (see Campbell, 2015).

Hansen's (2016) analysis of data from the late 60s found teachers to have a greater probability of over-rating, and a lower probability of under-rating, attractive pupils after statistically controlling for any association between attractiveness and academic ability. Shackleton and Campbell (2014) found little evidence of teacher judgement of reading and maths being influenced by pupils' waist circumference after controlling for any association between body shape and academic ability. These 2 studies serve to remind us that there are reports of teacher bias (or the lack thereof) in relation to

characteristics like students' physical attributes, students' personality and behaviour (see Urhahne & Wijnia, 2021 for a recent summary) – these characteristics have no apparent direct relationship with attainment, and one would be surprised if there were any demonstrable conditional achievement gaps by them in teacher assessment results.

Cornwell et al. (2013) focused on gender. They reported in the teacher assessment results they analysed an achievement gap conditional on test scores in favour of girls or against boys, and that the gap more or less vanished after factoring in ratings of the pupils' classroom behaviour given by their former teachers one or two years previously. The implication is that what underlay an apparent teacher bias in relation to gender was a bias in favour of good behaviour or against bad behaviour.

Five of the studies on the list above examined multiple variables and employed similar analysis methods to the studies reviewed in the 2 previous sections. Table 6 shows the conditional achievement gaps in teacher assessment that were found or not found in the various studies. Despite the differences noted above in the nature of the data used, the findings are highly similar to those summarised in the previous 2 sections. On gender, divergence in favour of girls or against boys was commonly found. On ethnicity, the findings were again mixed in that there were findings of divergence against as well as in favour of minority groups and findings of no systematic divergence. On socio-economic status, as in the England NCA studies, divergence in favour of the less disadvantaged or against the more disadvantaged was a common finding. In relation to SEN, divergence against SEN pupils or in favour of those without SEN was a common finding. On EAL, divergence against EAL pupils or in favour of non-EAL pupils was not a common finding, except in the subject of English.

Perkins et al. (2004) calculated correlations between grade point averages recorded on students' high school transcripts (which can be taken as results from teacher assessment) and their scores on the NAEP assessment, which were 0.49 for science and 0.53 for maths. They also presented the correlations for many subgroups of students. Unequal correlations can be taken as evidence of systematic divergence between results from the 2 assessment methods (for example, in science, 0.36 for black and Hispanic students, 0.48 for white students, 0.58 for Asian/Pacific Islander students), but we cannot tell from the report whether the differences in correlation were statistically significant.

In addition, it can be noted that in the studies grouped in this section, there were more explorations of the influences of teacher-, classroom- and school-level variables, not so much on the achievement gaps in relation to student characteristics in teacher assessment, but on the (relative) agreement between results from teacher and test-based assessments. For example, the New Zealand study found evidence that teacher assessment was lower for pupils in a higher-attaining classroom and in a higher-attaining school than for test-score-matched pupils in a lower-attaining classroom and in a lower-attaining school. The US literacy study reported that after conditioning on, among other things, standardised test scores, teacher assessment benefitted pupils in a higher-attaining classroom, those in a classroom with pupils with higher socioeconomic status, those with a less experienced teacher, and those in a school with pupils with lower socioeconomic status.

Table 6. *Conditional achievement gaps by student characteristics in teacher assessment in large-scale cohort studies or research projects*

	Campbell (2015) [UK Millennium Cohort Study; at age 7; English]	Campbell (2015) [UK Millennium Cohort Study; at age 7; maths]	Johansson et al. (2012) [Swedish PIRLS; 3rd and 4th graders; reading]	Martínez et al. (2009) [US ECLS-K; 3rd and 5th graders; maths]	Meissel et al. (2017) [New Zealand research project; 8- to 13-year-olds; reading and writing]	Ready & Wright (2011) [US ECLS-K; kindergarteners; literacy]
Gender	Female > Male	Male > Female	Female > Male	Female > Male	Female > Male	Female > Male
Ethnicity	White > Indian, Pakistani, Black Caribbean for Female No significant gap between White and Indian, Pakistani, Black Caribbean for Male No significant gap between White and Bangladeshi, Black African for Male and Female	White > Black Caribbean for Female Bangladeshi > White for Male No significant gap between White and Black Caribbean for Male No significant gap between White and Bangladeshi for Female No significant gap between White and Indian, Pakistani, Black African for Male and Female		Either No significant gap between Non-minority and Minority or Minority > Non-minority	European > Māori, Pasifika European > Other in reading No significant gap between European and Other in writing	White > Hispanic No significant gap between White and Black, Asian, Native American, Multiracial
Socio-economic status	Higher income > lower income	Higher income > lower income	Higher SES > Lower SES	Either No significant gap between Higher and Lower SES or Lower SES > Higher SES		Higher SES > Lower SES
SEN	NoSEN > SEN	NoSEN > SEN		NoSEN > SEN	NoSEN > SEN	

EAL	NotEAL > EAL for Male No significant gap for Female	No significant gap			NotEAL > EAL	No significant gap between NotEAL and EAL NotEAL > Asian EAL
Based on significance testing of relevant statistics in: (number of observations in analysis in brackets)	Table 8 (4997)	Table 9 (4985)	Figure 8 and accompanying text in PhD dissertation (11315)	Interpretation of conditional attainment gaps inferred from authors' interpretation of d values in Table 6 (10700 [3rd grade], 8600 [5th grade])	Table 4 (4771 [reading], 11765 [writing])	Spring Model 2 of Table 3 (9493)

Discussion

Following the cancellation of exams, GCSE, AS and A level grades of summer 2021 will be based on teacher judgements. To raise awareness of potential risks to the dependability of those judgements, we conducted a review of research evidence of systematic divergence between results from teacher and test-based assessments. Such evidence does not prove error or bias in teacher judgements, but it points to the possibility of bias in teacher assessment judgements given their evidently greater subjectivity and hence greater vulnerability to bias relative to test results.

We grouped the studies providing relevant research evidence into 3 types based on broad classification of their data source. The data sources have features that may impinge on the relevant findings' generalisability, both to our understanding of teacher assessment in general and to informing us of what to expect about the grades of summer 2021. Those features include the extent to which the teacher and test-based assessments measured (or were designed to measure) the same construct, how independently the two assessment methods operated in the respective system/study, how high the stakes of the assessments were for students, whether the students' attainment could have genuinely changed in the intervening time between the assessments, what the teacher assessment entailed and its status, and the quality of, and potential bias in, the comparator test-based assessment. Another feature we should add is the level of education that the evidence pertained to. With the exception of the evidence in the Israeli and the US HSTS/NAEP studies (and probably also the Norwegian and Spanish studies), all the evidence came from lower, and in many cases, much lower, levels of education than the ones with which we are presently most concerned. We should also be mindful of possible publication bias in educational research (see, for example, Torgerson, 2006): do we come across more reports of systematic divergence than of no divergence because findings of systematic divergence are more likely to get published?

The unique circumstance under which teacher judgements are called upon for summer 2021 means that the teacher assessment to be conducted has no exact parallel in the literature we have reviewed, and no study reviewed is unreservedly informative about summer 2021. We cannot conclude with certainty whether, or in what way, the teacher-assessed grades of summer 2021 will be biased relative to the counterfactual, would-have-been exam grades. Nevertheless, some repeated findings in the literature of systematic divergence between results from teacher and test-based assessments suggest possible biases that are worth drawing attention to.[9]

On gender, bias in favour of girls or against boys in teacher assessment results was more commonly found than no bias or bias in favour of boys or against girls. On ethnicity, there were findings of bias against as well as in favour of each minority group (relative to the majority group) and findings of no bias. On socio-economic status, bias in favour of the less disadvantaged or against the more disadvantaged

[9] Recall that we identified systematic divergence through statistically significant effects. As noted above, statistical significance is partly dependent on sample size. As large datasets were analysed in the studies we reviewed, some of the effects found, while statistically significant, had small effect sizes. Some researchers commented on relative effect sizes (for example, Gibbons & Chevalier, 2008), but one cannot find in this literature an effect size criterion for distinguishing between educationally significant and non-significant effects.

was a more common finding than no bias in UK-based studies. Bias against pupils with SEN or in favour of those without was found in nearly every analysis that included the SEN status variable. Bias against EAL pupils or in favour of non-EAL pupils was not a common finding, except in the subject of English.

How can these biases be explained? A few studies of NCA began with the assumption that biases in teacher assessment reflect teachers' differential expectations of students. Teacher expectations are widely known, or believed, to have a 'Pygmalion', or self-fulfilling prophecy effect on students: teachers' high expectations of students lead to better attainment in students and low expectations lead to worse attainment (see Jussim & Harber, 2005 for a meta-analysis of studies of the phenomenon). The Pygmalion effect, if robustly present, would lead to differential attainment for students with differing levels of expectation placed on them. As attainment should affect achievement at teacher and test-based assessments equally and the biases in teacher assessment we are presently concerned with are evidenced by achievement gaps conditional on test scores, they seem unconnected with the veracity of the Pygmalion effect.

Re-labelling biases as differential expectations does not take us very far. One may then question where differential expectations come from. One attempt to explain biases or differential expectations builds on the idea that categorisation is a fundamental cognitive ability. To organise or simplify our experiences and knowledge of the world, we put objects into categories on the basis of their shared features or similarities. Social categorisation is the process by which we categorise people into social groups along the lines of gender, ethnicity, social class and so on. and think of them as members of a social group rather than as individuals. In teachers, the natural process of social categorisation, coupled with primarily experiences in their own schools and probably also knowledge gained from the education system and the wider society, leads to the development of stereotypes, that is, generalised expectations and beliefs about the characteristics of particular groups of pupils. The stereotype of a group helps save time and effort in making judgements about individuals belonging to that group but can lead to erroneous judgements on at least some members of the group, because of its generalised nature and possible inaccuracy. Stereotyping can be the mechanism underlying teacher bias against low-attaining groups or in favour of high-attaining groups.

The notion of stereotyping sits less well with the less common but not exactly rare findings of teacher bias against high-attaining groups or in favour of low-attaining groups. It has been suggested that such biases reflect teachers' counter-stereotyping or compensatory grading.

In addition to the ideas of teacher expectation, stereotyping and counter-stereotyping, there are other accounts that do not so much explain as explain away particular biases. For example, we mentioned above Cornwell et al.'s (2013) demonstration that in their teacher assessment data what appeared to be a teacher bias in relation to gender could be reduced to a bias in favour of good behaviour or against bad behaviour. We note, however, that Burgess and Greaves (2013) argued that ethnic differences in classroom behaviour could not explain the bias in relation to ethnicity in their teacher assessment data. Another account, often discussed with reference to the bias in relation to SEN status, considers an extreme form of 'teaching to the test': teachers may teach to the test more with some students. A consequence of teaching to the test is students' possible over-achievement on the

test relative to their actual level of attainment, which in turn makes an accurate teacher rating of the actual level of attainment look like an under-rating. In other words, what appears to be a bias against a group in teacher assessment may be explained in terms of differences in test-specific learning.

Conclusion

By way of conclusion, we discuss the similarities and differences between the present review of teacher assessment and our review last year of teacher prediction. It should be borne in mind that we may not be comparing the 2 reviews on equal footing as last year's review covered mainly research on teacher judgement/prediction at GCSE and A level while for the present review, we managed to find research on teacher judgement/assessment mostly at lower levels of education.

The 2 reviews suggest that the relative agreement (as measured by similarity in the rank order of students) between results from teacher and test-based assessments is of a comparable level to the relative agreement between teacher prediction and actual achievement. There appears to be a higher level of absolute agreement between results from teacher and test-based assessment in NCA than between teacher prediction and students' actual achievement at GCSE and A level, but it should be noted that the teacher and test-based assessments in NCA did not operate independently. We saw in last year's review ample evidence of teachers' tendency towards over-prediction in grade prediction/forecast. We found little evidence of the equivalent (that is, teacher over-rating relative to test results) in NCA studies, but in 2 studies from abroad that provided rare analyses of absolute agreement, we saw that in case of disagreement between teacher and test grades, over-rating by teachers (relative to the test grades) was much more likely than under-rating, both in the Dutch system where teachers were aware of students' test results when finalising their judgements and in the Norwegian system where they were not.

On gender, the findings were mixed in last year's review on teacher prediction, and in the present review on teacher assessment, a slight female advantage or male disadvantage (relative to test results) is the more common finding.

On ethnicity, we saw in last year's review that in several combined-subject analyses, predictions, relative to actual achievement, were higher for black and Asian students than for white students. The equivalent in teacher assessment, that is, higher teacher ratings relative to test results, was reported occasionally for black students and commonly for Asian students, but there were also findings of the opposite and no difference. It is difficult to make a systematic comparison between the 2 reviews' findings on ethnicity because the teacher prediction analyses did not use the finer ethnicity classification used in some of the teacher assessment studies and there is no combined-subject analysis in the teacher assessment literature.

On socioeconomic status, we saw in last year's review that predictions, relative to actual achievement, were higher for the more disadvantaged than for the less disadvantaged. The opposite, that is, lower teacher ratings, relative to test results, was commonly found in the teacher assessment literature, if we consider only the UK-based studies.

On SEN and EAL status, the teacher prediction analyses had little to say. In the present review of teacher assessment, we have seen a common finding of lower teacher ratings, relative to test results, for students with SEN than for those without, but no evidence of teacher bias in relation to EAL status in any subject except for English.

References

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578868/2011-03-16-estimates-of-reliability-of-qualifications.pdf

Brookhart, S.M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20(1)*, 69-90. http://dx.doi.org/10.1080/0969594X.2012.703170

Burgess, S., & Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics, 31(3)*, 535-576. https://doi.org/10.1086/669340

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy, 44(3)*, 517-543. https://doi.org/10.1017/S0047279415000227

Connolly, P., Taylor, B., Francis, B., Archer, L., Hodgen, J., Mazenod, A., & Tereshchenk, A. (2019). The misallocation of students to academic sets in maths: a study of secondary schools in England. *British Educational Research Journal, 45(4)*, 873–897. https://doi.org/10.1002/berj.3530

Cornwell, C., Mustard, D.B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school. *Journal of Human Resources, 48(1)*, 236-264. https://doi.org/10.3368/jhr.48.1.236

Department for Children, Schools and Families. (2009). *Children with special educational needs 2009: an analysis*. Nottingham: DCSF Publications. https://dera.ioe.ac.uk/9446/1/Main.pdf

Department for Education (2019). *Timpson review of school exclusion: technical note*. London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/799910/Technical_note.pdf

Durant, D. (2013). A comparative analysis of Key Stage tests and teacher assessments. Paper presented to the British Educational Research Association Annual Conference (September 2003) at Heriot-Watt University, Edinburgh. http://www.leeds.ac.uk/educol/documents/00003153.doc

Falch, T., & Naper, L.R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review, 36*, 12-25. https://doi.org/10.1016/j.econedurev.2013.05.002

Feron, E., Schils, T., ter Weel, B. (2016). Does the teacher beat the test? The value of the teacher's assessment in predicting student ability. *De Economist, 164*, 391–418. https://doi.org/10.1007/s10645-016-9278-z (also a January 2015 working paper, accessed on 24 March 2021 at http://www.academischewerkplaatsonderwijs.nl/files/3414/2070/3590/WP_Feronea_2015.pdf)

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 (also a December 2007 working paper, accessed on 14 March 2021 at https://personal.lse.ac.uk/gibbons/papers/Teacher%20Assessments%20December%202007.pdf)

Hansen, K. (2016). The relationship between teacher perceptions of pupil attractiveness and academic ability. *British Educational Research Journal, 42(3)*, 376–398. https://doi.org/10.1002/berj.3227

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

Johansson, S., Myrberg, E., & Rosén, M. (2012). Teachers and tests: assessing pupils' reading achievement in primary schools. *Educational Research and Evaluation,18(8)*, 693–711. https://doi.org/10.1080/13803611.2012.718491 (Also Johansson's 2013 PhD dissertation, accessed on 26 March 2021 at https://gupea.ub.gu.se/handle/2077/32012)

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28(1)*, 91-105. https://doi.org/10.1080/02671522.2012.754229

Jussim, L., & Harber, K.D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9(2)*, 131-155. https://doi.org/10.1207/s15327957pspr0902_3

Kaufmann, E. (2020). How accurately do teachers judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*, 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009

Lee, M.W., Stringer, N., & Zanini, N. (2020). *Student-level equalities analyses for GCSE and A level: summer 2020*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/938869/6713_Student-level_equalities_analyses_for_GCSE_and_A_level.pdf

Lee, M.W., & Walter, M. (2020). *Equality impact assessment: literature review*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879605/Equality_impact_assessment_literature_review_15_April_2020.pdf

Lindahl, E. (2016). Are teacher assessments biased? Evidence from Sweden. *Education Economics, 24(2)*, 224-238. https://doi.org/10.1080/09645292.2015.1014882

Malouff, J.M., & Thorsteinsson, E.B. (2016). Bias in grading: a meta-analysis of experimental research findings. *Australian Journal of Education, 60(3)*, 245–256. https://doi.org/10.1177/0004944116664618

Marcenaro-Gutiérrez, 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

Martínez, J.F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment, 14(2)*, 78–102. https://doi.org/10.1080/10627190903039429

Perkins, R., Kleiner, B., Roey, S., & Brown, J. (2004). *The High School Transcript Study: a decade of change in curricula and achievement, 1990-2000*. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2004455

Pinot de Moira, A. (2020). *The impact of coursework on attainment dependent on student characteristics: a study based on GCSE and A level outcomes between 2004 and 2017*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/896472/The_impact_of_coursework_on_attainment_dependent_on_student_characteristics.pdf

Plewis, I. (1997). Inferences about teacher expectations from national assessment at key stage one. *British Journal of Educational Psychology, 67(2)*, 235-247. https://doi.org/10.1111/j.2044-8279.1997.tb01240.x

Rangvid, B.S. (2015). Systematic differences across evaluation schemes and educational choice. *Economics of Education Review, 48*, 41-55. https://doi.org/10.1016/j.econedurev.2015.05.003

Ready, D.D., & Wright, D.L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: the role of child background and classroom context. *American Educational Research Journal, 48(2)*, 335-360. https://doi.org/10.3102/0002831210374874

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)*, 141–160. https://doi.org/10.1080/0141192012003710

Rhead, S., Black, B., & Pinot de Moira, A. (2016). *Marking consistency metrics*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/att

achment_data/file/681625/Marking_consistency_metrics_-_November_2016.pdf

Rhead, S., Black, B., & Pinot de Moira, A. (2018). *Marking consistency metrics: an update*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Rimfeld, K., Malanchini, M., Hannigan, L.J., Dale, P.S., Allen, R., Hart, S.A., & Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable and heritable as standardized test scores. *Journal of Child Psychology and Psychiatry, 60(12)*, 1278–1288. https://doi.org/10.1111/jcpp.13070

Shackleton, N., & Campbell, T. (2014). Are teachers' judgements of pupils' ability influenced by body shape? *International Journal of Obesity, 38(4)*, 520–524. https://doi.org/10.1038/ijo.2013.210

Strand, S. (2012). The White British-Black Caribbean achievement gap: tests, tiers and teacher expectations. *British Educational Research Journal, 38(1)*, 75-101. http://doi.org/10.1080/01411926.2010.526702

Strand, S., & Hessel, A. (2018). *English as an additional language, proficiency in English and pupils' educational achievement: an analysis of local authority data*. Cambridge: The Bell Foundation. https://mk0bellfoundatiw1chu.kinstacdn.com/app/uploads/2018/10/EAL-PIE-and-Educational-Achievement-Report-2018-FV.pdf

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology, 104(3)*, 743–762. https://doi.org/10.1037/a0027627

Thomas, S., Smees, R., Madaus, G.F., 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

Torgerson, C.J. (2006). Publication bias: the Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54(1)*, 89-102. https://doi.org/10.1111/j.1467-8527.2006.00332.x

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Wheadon, C., & Stockford, I. (2010). *Classification accuracy and consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009*. Coventry: Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/578862/2011-03-16-aqa-classification-accuracy-and-consistency-in-gcse-and-a-levels.pdf

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

May 2021

Ofqual/21/6781