British Journal of Psychology Research
Vol.8, No.1, pp. 26-35, February 2020
Published by ***ECRTD-UK***
Print ISSN: 2055-0863(Print), Online ISSN: 2055-0871(Online)

# EXAMINING THE ADEQUACY OF PAST PERFORMANCES AND TEACHER JUDGMENTS TO ESTIMATE GCSE AND A-LEVEL GRADES DURING THE PANDEMIC CRISIS

**Matthew Rudd, PhD**
English Language Centre, Mahanakorn University of Technology, Bangkok 10530, Thailand

**ABSTRACT**: *On the 17th March, 2020, the British government announced that due to school closures during the height of the pandemic, GCSE and A-Level grades would be determined on the basis of teachers*' *assessments and evidence of students*' *hitherto attainments*. *England*'*s Education Secretary, Gavin Williamson, emphasised that* "*grades awarded this summer will accurately reflect students' abilities and will be as valid this year as any other*".*To test this assertion, this quantitative research study independently examined the predictive accuracy of using past performances and teacher judgments to anticipate students*' *final grade outcomes among 84 third-year university students at a private university located in the outskirts of Bangkok*. *After assessments were formally graded, the mean end-of-year English score for 2018-2019 was calculated at 59.94%, which was significantly different to the average standard attained previously in 2017-2018* (*54.36%; sig, p <0.05*). *Nonetheless, in disagreement with prior literature, teacher judgments proved to be statistically reliable* (*57.98%, not sig, p <0.05*). *Further implications, research recommendations and policy considerations are also discussed in this paper*.

**KEYWORDS**: performance, teacher, judgment, accuracy, attainment

## INTRODUCTION

As an urgent response measure to the Covid-19 pandemic wreaking havoc across the nation, the UK government swiftly announced the closure of schools commencing Friday 20th March 2020, as reported on BBC News. Despite the lockdown measures, one month later, England's School Standards Minister Nick Gibb assured that Grades for cancelled GCSEs and A-level exams will still be published on the conventional dates; A-levels on the 13th of August and GCSEs on the 20th; adding that grades would be determined by teachers' estimations, (reported on BBC NEWS, 20th April).

Mary Bousted, joint leader of the National Education Union offered strong assurances stating that "grades won't be based on mock exam results, or any other *single* piece of evidence alone". The collection of evidence comprises teachers' predictions for A-levels, AS-levels and GCSEs, prior exam results, in-class tests, standard of homework, formally assessed coursework, and, mock exams. Sally Collier, chief executive of Ofqual's (UK's qualifications watchdog) insisted that "Our overriding aim in this is to be fair to students this summer and to make sure you are not disadvantaged,". The central focus of this paper is to examine the reliability and fairness of resorting to past performances and teacher judgments as an alternative means of awarding final grades.

Past performance

Given the dissimilar nature of academic subjects and their respective assessment criteria, information collated deriving from past assessments may construct predictive models without considering students' current activity data (Hellas el at., 2018); which is not available to teachers during the prolonged closure of schools. Students may feel that retrospective analysis of past performances to hypothetically determine would-be future grades condemns them to failure (Marbouti et al., 2016), removing their ability to exert influence in their current life circumstances. Given the specific course demands of academic subjects in higher education, the implementation of admission tests based on content-valid methods to enhance the sorting and selection of entry-level candidates has become universally mainstream (Schmitt, 2012). In the UK however, the selection process of university applicants depends on UCAS scores, which involves converting letter grades of individual A-level subjects into a numerical value. Thus, admissions are based on whether a candidate's cumulative UCAS score meets that of the university's entry requirements. Kuncel, Hezlett, and Ones (2001) concluded through meta-analysis across multiple disciplines that specific subject tests administered prior to entry most adequately predicted university students' GPAs at under-graduate and post-graduate levels, than did the general standard of high school achievements; further reviews (Kuncel and Hezlett, 2007) lent further support to their earlier findings.

**Teacher judgements (JA)**

Teacher judgment accuracy has been described by Artlet & Raush (2014) as an ability to judge student characteristics correctly, and, is part of a broader set of skills known as "diagnostic competence", referring to a teacher's ability to accurately predict task demands that students are capable of (Anders, Brunner, & Krauss, 2011). Students categorised as more capable are increasingly more likely to benefit from higher quality learning opportunities (Clark & Peterson, 1986), and teachers' assessments of students are key decisional determinants for student placement in programs, ability groups, grade retention, and, students' future academic itineraries (Francis et al., 2016). Hoge & Coladarci (1989) and Südkamp, Kaiser, & Möller (2012) categorise judgments on academic performance as methodological variables, whereas moderating variables are linked to external factors, such as the gender, age and culture of the teacher / student. Hoge and Coladarci (1989) pointed out that the accuracy of teacher judgments is susceptible to variability and inconsistency, as brought to light in their review of 16 published articles which highlighting that r coefficients fluctuated exponentially between .28 and .92.

## LITERATURE REVIEW

The review of literature below discusses an compilation of studies conducted around the globe testing the predictive utility of past performances and teacher ratings.

*Past performance: psychology students (Netherlands)*

This paper examined the validity of predicting students' first-year academic performance through university entrance testing, (referred to as trial-studying test), and, judging by applicants' prior high school attainments. The study took place during the academic year 2013-2014 at a Dutch university and comprised 652 applicants (mean age being 20 years), who were applying for an undergraduate psychology program. Students were required to study two chapters of course material; the first chapter was about research methodology and the second covered general psychological theories. Students were awarded grades on a scale of one (the lowest) to ten (the highest), and 6 was the threshold for a pass mark. Students then sat an exam consisting of 40 questions in multiple-choice style response format.
The results showed that the university admission test proved to be the most accurate predictor of end-of-year academic performance ($r$ = .56) for the first year; yielding predictive validities similar to those of prior educational performance ($r$ = .52). Nevertheless, the latter measure only proved to be reliable with students coming from a Dutch educational background, as half of the students were from overseas, and thus educational backgrounds were heterogeneous which complicated benchmarking. As a result, the trial-studying entrance test proved to be the most solid predictor of end-of-year grade point averages.

While the results from this study lend support to the fact that past performance deriving from relevant domains is a reliable indicator of future performances, universities in the UK do not conduct entrance examinations, and thus, under the current critical circumstances, sorting procedures are currently constrained to proxy tools to judge the progress of students' applications. The study reviewed above also put forward the claim that prior achievements of students from *standardised* education backgrounds proved to correlate with successive end-of-first-year achievements in higher education, despite involving different assessment criteria and very dissimilar learning environments to those in secondary education. This issue will re-emerge for further discussion later in this article.

*Past performance: dentistry students (New Zealand)*
In line with the above research, an article published in the European Journal of Dental Education (Vol 22, issue 2, 2018) also tested the predictive strength of prior performances among a sample 116 oral health students, which also analysed variables such as age, gender, ethnicity, level of prior education, institute of prior education, and, work experience. The aim was to predict students' academic performance in the discipline-specific bioscience paper at the Auckland University of Technology from 2011 to 2014 through use of multivariable regression analysis and Pearson's Correlation Coefficient. The results determined that prior academic achievement was the only statistically significant predictor of academic performance in the discipline-specific bioscience paper, ($r$ = 0.641). Interestingly, the results confirmed the predictive strength of past performances in a discipline-specific domain, but discounted the effectiveness of assessing prior educational background as a means of anticipating future academic behaviour. Similar to the first study discussed in this section, the reliance on Pearson's correlation coefficient, however, becomes questionable. This statistical instrument is better used for detecting patterns and relationships,

rather than distinguishing the level of significance between two variables; a strong correlation can positively link two performance measures of significantly dissimilar standards of achievement.

*Teacher judgments: in literacy (Germany)*

Kaiser et al. (2013) examined the reciprocal relationship between teacher judgments and students' successive reading achievements. The students took part in a German literacy development project and sat a reading test from the German PIRLS study. Teacher judgments were collated in the form of two four-point response scales to gauge students' literacy competence. The study revealed a moderate correlation of r = 0.49 between teacher judgments and students' corresponding achievements. Nonetheless, the relatively weak correlation could be due to the fact that teachers were not made aware of the criteria of achievement against which their judgments were compared; despite close familiarity with the students whose scores they were predicting. This arguably highlights a flaw within this paradigm of research, given that comparative criteria was not made clear to the teachers, as well as the fact that estimations were scored on a very limited continuum of 1 – 4; likely clouding judgment. The current study will select three teachers who are familiar with their students, the curriculum, the corresponding assessments, and are able to indicate their judgments on a continuum of 1-100 (%).

*Teacher judgments: young children's literacy (USA)*

The wide-ranging study by Ready and Wright (2011) compiled a collection of data from the Early Childhood Longitudinal Survey, tracking the progress of 10,000 children's literacy achievements in the United States ranging from kindergarten to fifth grade. Children's attainments were consequently compared with teacher's prior predictions. Ready and Wright (2011) discovered that teachers in urban schools tended to underestimate students' ability, whereas teachers in small schools had the tendency to overestimate students' literacy skills.The findings uncovered systematic bias trends in teachers' judgments, who had reportedly overestimated girls' attainment potential, while underestimating students from Black, Hispanic and Asian backgrounds. The authors also concluded that negative bias evidently affected the accuracy of teacher judgments, especially for students from a low socioeconomic background, raising concerns that an emerging systematic bias among teachers risked further marginalisation of learners. Such partiality in judgments may risk interfering with teachers' estimations in the current state of affairs across England, Scotland, Wales and Northern Ireland during the pandemic crisis, despite assurances of externalmoderation.

**Research objective**

This research paper examined the validity of two independent predictive instruments which included prior performances and teachers' estimations as an alternative means of awarding final grades during the unprecedented period of prolonged school closures. Data from both measures were subsequently paired against students' final grade outcomes in English via t-test calculations. In line with Ready and Wright's (2011) revelations, further meta-analyses were carried out to detect potential bias in teacher judgments in relation to gender. Unfortunately, exploring further

intervening bias as a function of socioeconomic backgrounds and ethnicity fell out of the scope of the current study.

**Hypothesis**

Based on the review of related studies that attest to the predictive reliability of past performances, and, the body of research that questions the accuracy teacher judgments; past performances are expected to offer greater reliability.

## METHODOLOGY

Following executive approval, this study was authorised to take place during academic year of 2018-19 at a private university situated in the surrounding areas of Bangkok. Senior management also kindly provided all data pertaining to official grades.

The central focus of the study was to examine the predictive efficacy of past achievements and teacher ratings to predict students' end-of-year test scores in English.

RESEARCH DESIGN

The methods used for obtaining the data for past performance and teacher ratings are outlined below:

*Past performances*: The university registrar provided final grade outcomes from the previous English courses taken in the academic year of 2017-2018. These prior attainments were contrasted with the students' end-of-year performances for the academic year of 2018-2019. Itemised data concerning speaking assessments and individual examination scores were not available as students' end-of-year results for are archived as one overall percentage score.

*Teacher judgments*: Three teachers, including the researcher, partook in this experiment. All teachers were highly aware of students' strengths, weaknesses, academic interests, and future career plans. Prior to the mid-term tests, teachers were asked to independently predict the overall grades of their students on a continuum of 1-*100 (90 = A, 80 = B+, 70 = B, 60 = C+, 50 = C, 40 = D, <39 = F)*. Estimations were subsequently compared with final grades. Comparisons according to class were not disclosed, as the concept of this research design was not to cross-compare the teachers' individual judgment efficacy.

**Participants**

The sample population consisted of 84 third year university students (females = 51; males = 33) from a private higher education institute located on the outskirts of Bangkok. The students study English as a separate subject and were from two separate faculties; business administration (n = 58) and engineering (n = 26)

**Research Procedure**

The assessments for English language courses comprise three parts: mid-term tests (weighted at 20%), speaking tests (30%) and end-term examinations (50%); both written examinations consist of reading, writing and grammar based sections. Teacher judgments and past attainments were then compared with the final grade outcomes using t-tests.

**FINDINGS / RESULTS**

Statistical analyses looked to cross-examine the accuracy of prior English language attainments and teacher judgments to predict students end-of-year grades in English.

Table 1: Past performances and teacher judgments vis-à-vis end-of-year English language scores

| **English** | **Measure** | **M** | **P** | **SD** | **Result <.05** | **R** |
|---|---|---|---|---|---|---|
| Mean of N 59.94% (SD: 20.21) | Past perf. | 54.47% | 0.026 | 16.59 | Sig | 0.842 |
| | T. Judgments | 57.98% | 0.248 | 16.74 | Not Sig | 0.707 |

*Note: N = 84*

As seen in Table 1, teacher judgments proved to be significantly more reliable than past performances. This is based on the observation that there was no significant difference between the grades teachers predicted, and, the grades students went on to achieve (TJ: 57.98% vs. 59.94%; $p = 0.026$). This disproves the hypothesis set out in this paper, which was formulated on the outcome of Hoge and Coladarci's literature review and Kaiser et al.'s (2013) study that established a weak and unreliable correlation between teacher judgments and subsequent performance. The findings in Table 1 also refute the reliability of past performance to predict future achievements. Prior literature (Neisen, Meijer & Terneiro, 2016; and, Khareedi, 2018) both claimed that past performances can firmly predict future ones, a claim that the results displayed in Table 1 do not support.

As noted in Table 1 above, t-test results revealed that students' end-of-year attainments in English were significantly different to their past performances (past performance: 54.47% vs. 59.94%; $p = 0.026$), insinuating that past achievements do not serve as a reliable projector of future academic behaviour; despite inferring stronger correlations (past performance: $r = 0.842$ vs. TJ: $r = 0.707$). The closer correlation noted among past performance may be owed to the fact that students' end-of-year improvements were incrementally similar, and assumedly this form of calculation fails to highlight the level of difference between prior and current achievements. Under the university's grading criteria in this setting, students scoring 54.47% would have been awarded a C grade, while those attaining 59.94% (rounded up to 60%) would have achieved a C+, which could have a notable impact on one's grade point average.

| **Final English Grades** | | **Past Performance** | | **Teacher Judgments (TJ)** | |
|---|---|---|---|---|---|
| Females | Males | Females | Males | Females | Males |
| 60.2% | 59.55% | 54.76% | 52.94% | 61.96% | 51.82% |
| (M of N: 59.94%) | | (M of N: 54.47%) | | (M of N: 57.98%) | |
| p = 0.472 | p = 0.462 | p = 0.382 | p = 0.333 | p = 0.084 | p = 0.041 |
| not sig | not sig | not sig | not sig | not sig | sig |

Table 2: the gender variable in past performances and teacher judgments

*Note: N = 84; females = 51; males = 33*

As reflected in the results displayed in Table 2, females narrowly outperformed their male counterparts (females: 60.2% vs. males: 59.55%), notwithstanding male students narrowed the achievement gap, and their rate of improvement almost reached statistical significance:
(past performance for males: 52.94% vs. end-of-term: 59.55%; p = .0611, not sig <.05);
(past performance for females: 54.76 % vs. end-of-term: 60.2%; p = .0829, not sig <.05).
This statistical phenomenon emerges despite significant improvements being noted across the collective population of participants (see Table 1).
In relation to teacher judgments, the results displayed in Table 2 lend firm support to Ready and Wright's (2011) observation that teachers are more generous with their estimations of female students, and very much underestimated the capabilities of the male students (TJ for females: 61.96% vs. final grades: 60.2%; p = .313, not sig <.05); (TJ for males: 51.82% vs. final grades: 59.55%; p = .0494, sig <.05). In the case of the latter, teachers mistakenly conferred the *average* male student a C grade, instead of a C+.

The micro analyses in Tables 1 and 2 highlight four implications:
**(1)** Prior attainments exclude the possibility of future improvements.
**(2)** Teacher judgements proved to be more accurate (especially for female students).
**(3)** Teacher judgments underestimated the potential of male students.
**(4)** T-tests appear to be more informative than Pearson's coefficient in this context.

**Implication to Research and Practice**
In light of the statistics analysed, the following discussion will seek to clarify a number of complex issues, such as the validity of the hypotheses, the limitations associated with this study and suggestions for future research interventions. Despite Hoge and Coladarci's (1989) claim that teacher judgments are variable and unreliable, and, Kaiser et al.'s (2013) research findings implying that teacher judgments were moderately effective, the methodology of this paper was unique in the sense that the collaborating teachers in this study reported their findings directly to the researcher. The authors of the papers reviewed in literature did not personally implement

their own research experiments, rather relied on proxy institutions, whose teachers also were not always familiar with the judgment criteria. Moreover, Hoge and Coladarci (1989) redacted an essay of findings from existing studies, all of which were published prior to 1989; and with changing times, reference to more current research may be more suitable; prompting the current line of inquiry.

The main constraint within the context of the current study was the impossibility to investigate potential bias of teacher judgments in the context of other moderating variables such as ethnicity and socioeconomic background. This is due to the uniform demographics of Thai society, and most students in this study were from working class families. Nevertheless, this paper confirmed a clear bias tendency in favour of female students, which followed the same vein as Ready and Wright's (2011) assertions, and by extension, leads to the possible assumption that such extrapolations of bias judgment behaviour may also operate in teachers' estimations of students belonging to minority ethnic groups and those from unfavourable socioeconomic backgrounds.

## CONCLUSION

It was reported on BBC News (April 3rd, 2020) that Paul Whiteman, the leader of the National Association of Head Teachers, rightly admitted that there was no "perfect solution"; assuring that the recently announced approach to decide A-level and GCSE grades was "pragmatic and the fairest" during such exceptional circumstances.

Although pragmatic in the short term, the results in this paper did not support the notion of fairness. Prior performances failed to accurately project end-of-year grades, and teacher judgments, while generally accurate in the context of this particular experiment, did prove to be subjectively influenced by bias; as manifested in the disproportionate levels of confidence placed in female students relative to their male counterparts.

Further proposals have also been put forth for exams to be taken in the autumn despite the likely disruption to academic calendars. The findings in this paper questioned the effectiveness of referring to past performances to accurately predict future accomplishments, and confirmed that teacher judgments are prone to bias; thus, delaying exams until the autumn may be the most suitable course of action, untimeliness notwithstanding.

### Future Research

Given the rapidity at which viral outbreaks can become global pandemics, further research of assessing the reliability of past performances and teacher judgments should be a matter of national interest. Aside from trialling these variables across larger populations of students, in different educational settings, in different countries, and, across a range of subjects; investigating potential bias according to socioeconomic background and ethnicity deserve close investigative attention. An additional means of predicting students' grade outcomes under more regular circumstances could come in the form of self-efficacy scales. Self-efficacy refers to people's beliefs about their capabilities to produce designated levels of performance (Bandura 1994) and peoples' achievements can be predicted most accurately by measuring beliefs of personal efficacy, and have proven to be a better predictor of behaviour than past performances (Bandura, 1977; 1986). This

due to the fact that efficacy beliefs affect self-motivation and action through their impact on goals and aspirations (Bandura, 2009). Graham and Wiener (1996) added that the beliefs students create, develop, and hold to be true about themselves are vital forces in their success or failure in school; and, efficacy beliefs govern the level and intensity of the effort invested in endeavours (Pajares, 2003).

Schools could regularly contemplate paying close attention to students' perceived levels of self-efficacy across all subjects on a regular basis to identify high and low confidence areas to improve implementation of the curriculum and to address individual needs. Moreover, with efficacy information at hand, schools across the UK would have had an additional predictive tool at their disposal to help guide their predictions.

## REFERENCES

Anders, Y., Brunner, M., & Krauss, K. (2011). Diagnostic skills of mathematics teachers and the performance of their students. Psychology in Education, 3 (2), 175-193.

Artlet, C., & Raush, T. (2014). Accuracy of teacher judgments. When and for what reasons? In Krolak-Schwerdt, S., Glock, S., & Böhmer, M. (Eds.), Teachers' professional development: Assessment, training, and learning (pp. 27-44). Rotterdam, The Netherlands: Sense.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), Encyclopedia of human behavior (Vol. 4, pp. 71-81). New York: Academic Press. (Reprinted in H. Friedman [Ed.], Encyclopedia of mental health. San Diego: Academic Press, 1998).

Bandura, A. (1997). Self-efficacy: The exercise of control. New York: Freeman.

Bandura, A. (2009). Cultivate Self - efficacy for Personal and Organizational Effectiveness. In E.A Lock (E.D)., Handbook of principles of organization behavior. (2nd Ed). pp. 179-201. New York: Wiley

BBC News, (2020). Coronavirus: Teachers to estimate grades after exams cancelled. Reported on 20th March 2020 in Family and Education section.

BBC News, (2020). Coronavirus Teachers to grade students for cancelled exams. Reported on 3rd April 2020 in Family and Education section.
Clark, C.M., Peterson, P.L,. (1986). Teachers' thought processes.
M.C. Wittrock (Ed.), Third handbook of research on teaching, Macmillan, New York, NY (1986), pp. 255-296

Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B ., Travers, M.C. (2016). Exploring the relative lack of impact of research on 'ability grouping' in England: A discourse analytic account. Cambridge Journal of Education (2016), pp. 1-17.

Graham, S., & Weiner, B. (1996). Theories and principles of motivation. In D. C. Berliner & R. C. Calfee (Eds.), Handbook of educational psychology (pp. 63-84). New York: Simon & Schuster Macmillan.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... Liao,S. N. (2018). Predicting Academic Performance: A Systematic Literature Review. In Proceedings

Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (pp. 175-199). (ITiCSE 2018 Companion). New York, NY, USA.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. Review of Educational Research, 59 (3), 297-313.

Kaiser, J. Retelsdorf, J. Südkamp, A. Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. Learning and Instruction 28 (2013), pp. 73-84,

Khareedi, R. (2018) Predictors of academic performance in the discipline specific bioscience paper: a retrospective qualitative study European Journal of Dental Education 22, issue 2, 2018)

Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Comput. 103, 1–15.

Niessen, S., Meijer, R., Tendeiro, J. (2016). Predicting Performance in Higher Education Using Prximal Predictors. PLOS ONE (11) 4.

Pajares, F. (2003). Self-efficacy beliefs, Motivation and Achievement in Writing: A review of the literature. Reading & Writing Quarterly, (19) pp. 139 – 158.

Schmitt N. (2012). Development of Rationale and Measures of Noncognitive College Student Potential. Educ Psychol. 47 18–29.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. Journal of Educational Psychology, 104 (3), pp. 743-762.

Ready, D., Wright, W. (2011).Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom contextAmerican Educational Research Journal, 48 (2) pp. 335-360.