BMJ Open

Differential attainment among 1,512 medical students at four Scottish medical schools

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046056
Article Type:	Original research
Date Submitted by the Author:	22-Oct-2020

Complete List of Authors:	Hope, David; University of Edinburgh, Medical Education Unit Dewar, Avril; University of Edinburgh, Medical Education Unit Hothersall, Eleanor; University of Dundee, Medical Research Institute Leach , John ; Queen Elizabeth University Hospital, Institute of Neurological sciences Cameron, Isobel; University of Aberdeen, Applied Health Sciences (Mental Health) Jaap, Alan; The University of Edinburgh Edinburgh Medical School

Keywords:	MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS, AUDIT, EDUCATION & TRAINING (see Medical Education & Training), HEALTH SERVICES ADMINISTRATION & MANAGEMENT

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Differential attainment among 1,512 medical students at four Scottish medical schools

A retrospective analysis of change in attainment gaps during medical school study

David Hope[1], Avril Dewar[2], Eleanor J Hothersall[3], John Paul Leach[4], Isobel M Cameron[5] and Alan Jaap[6]

[1]Senior Lecturer in Medical Education, University of Edinburgh: david.hope@ed.ac.uk

[2]Fellow in Medical Education, University of Edinburgh: avril.dewar@ed.ac.uk

[3]Head of Undergraduate Medicine, University of Dundee: e.hothersall@dundee.ac.uk

[4]Head of Undergraduate Medicine, University of Glasgow, John.Leach@glasgow.ac.uk

[5]Senior Lecturer and Assessment Lead, University of Aberdeen, i.m.cameron@abdn.ac.uk

[6]Deputy Director of Teaching and Assessment Lead, University of Edinburgh, alan.jaap@ed.ac.uk

Address for Correspondence

Dr. David Hope

Medical Education Unit

The Chancellor's Building

The University of Edinburgh

College of Medicine and Veterinary Medicine

49 Little France Crescent, Edinburgh

EH16 4SB

Scotland, United Kingdom

david.hope@ed.ac.uk

Contributorship statement

Dr Cameron, Dr Hothersall, and Prof Leach were each responsible for sourcing data, describing the context, and exploring the results in their institutions. Avril Dewar was responsible sourcing data at her institution then collating all the data and running the initial analyses. Dr Hope organised the project, designed the analyses, was primarily responsible for writing the paper. Dr Hope is the guarantor for the content. Dr Jaap acted as supervisor for all the project work and reviewed the analyses.

All authors have separately reviewed the manuscript and provided input in developing the final analyses and paper.

Copyright

Competing interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors had financial support from the Scottish Medical Education Research Consortium (SMERC) for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Transparency statement

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned have been explained.

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Dissemination declaration

Results will be disseminated to the representatives of the study populations (medical student groups).

Role of the funding source

The Scottish Medical Education Research Consortium (SMERC) provided funding to allow the research project to take place. The funding was used to pay for administrator and researcher time to collate and analyse the data. The funder had no direct input into the analyses chosen or the reporting of the results. The researchers were independent from the funder, and all researchers had access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Data sharing statement

Due to the sensitivity of the dataset – including confidential information on student demographics and assessment scores – we are unable to share raw data.

Word count

2,965

Abstract

Objective: To measure Differential Attainment (DA) among Scottish medical students and to explore whether attainment gaps increase or decrease during medical school.

Design: A retrospective analysis of undergraduate medical student performance on written assessment, measured at the start and end of medical school.

Setting: Four Scottish medical schools (Universities of Aberdeen, Dundee, Edinburgh, and Glasgow).

Participants: 1,512 medical students who attempted (but did not necessarily pass) final written assessment.

Main outcome measures: The study modelled the change in attainment gap during medical school for four student demographic categories (white/non-white, international/Scottish domiciled, male/female and with/without a known disability) to test whether the attainment gap grew, shrank, or remained stable during medical school. Separately, the study modelled the expected vs. actual frequency of different demographic groups in the top and bottom decile of the cohort.

Results: The attainment gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). International, non-white, and male students outranked their comparison group at the start of medical school but fell behind by final assessment. No significant differences were observed for disability status. Students with a known disability, Scottish students and non-white students were over-represented in the bottom decile and under-represented in the top decile.

Conclusions: The tendency for attainment gaps to grow during undergraduate medical education suggests that educational factors at medical schools may – however inadvertently – contribute to Differential Attainment. It is of critical importance that medical schools investigate attainment gaps within their cohorts and explore potential underlying causes.

Article Summary

Strengths and limitations of this study

- This the largest study to date investigating longitudinal attainment gaps within undergraduate medical education
- By evaluating Differential Attainment longitudinally, the study tests whether attainment gaps are due to pre-existing group differences or emerge during medical school
- The study methodology shows attainment gaps are not static – gaps change in magnitude and even reverse direction during medical school
- Although the study offers insights into the scope of attainment gaps, it cannot explain the mechanisms behind them

Introduction

Promoting fairness in assessment is a key priority. Success in medicine should be determined by ability rather than background characteristics like ethnicity, sex, or socioeconomic status. (1) There is an increasing emphasis on educational processes being "fair" to candidates of diverse backgrounds: besides the legal and regulatory requirements (2) there is growing acceptance that evaluating fairness should be a routine part of test construction and assessment. (3)

Despite this, background characteristics continue to predict performance in medical education and training. The tendency for candidate performance to vary by group membership is usually termed Differential Attainment (DA), and it influences every stage of medical education. The likelihood of entering medical school is significantly affected by a range of background characteristics and many selection methods are also influenced by these variables. (4) Performance on measures of success at or just beyond graduation show a similar pattern (5) and, for example, ethnically white UK graduates out-perform non-white UK graduates in postgraduate examinations with typically moderate (d = 0.22) effects. (6) After graduation, ethnically non-white and female doctors experience lower levels of success on a range of professional and educational outcomes. (7,8)

Such compelling evidence has led to calls to establish the mechanisms of DA, but this is challenging. Many historical assumptions – such as the idea that examiners are biased against some candidate groups – remain commonly cited despite evidence to the contrary. (9,10) Examiner bias does not appear to explain DA in postgraduate clinical examinations (11) or written assessment. (12) Qualitative research has emphasised a range of possible factors that can contribute to DA, including trust between trainers and trainees and the process by which those in difficulty are identified and referred to support networks. (13–15) Other research has suggested that unconscious biases may alter training pathways or assessment in the workplace. (8,16)

As a result, evidence for the existence of DA is very strong but we have so far only a limited understanding of the mechanisms by which it operates or even whether DA increases or decreases with time spent in medical education. Compounding this, while a great deal of research has been carried out on access to medical school and postgraduate assessment, relatively little work has evaluated DA on assessment *during* medical school. In a large meta-analysis eleven of fourteen published studies examining undergraduate medical education used a single site and two of the remaining studies used only two sites. (6) Combined with the tendency to monitor attainment at only a single time point (typically finals) we know little of whether DA is of similar magnitude for different medical schools or remains stable during medical school.

This is an obvious limitation given the role of medical schools in providing the foundation of medical education and training. Due to the diversity of intakes, assessment choices, curriculum design and performance on postgraduate assessment (17,18) investigating DA at medical schools may help in several ways. By comparing different institutions, the effect of different recruitment strategies, curriculum types and policies on fairness in medical education can be explored. If the magnitude of DA is highly variable across institutions, it argues for a relatively larger role in medical school policy in creating DA. If DA remains consistent despite varying institutional contexts, it argues DA is either explained by factors outside of medical school control or that no current approaches are identifiably superior or inferior. By examining the data longitudinally, it becomes possible to explore whether DA increases or decreases over time. If DA is present from the earliest part of medical education, this suggests different mechanisms than if DA is minimally present at the beginning but then grows with time. Such work can therefore significantly improve medical education and support a fairer experience for doctors.

In this study, we use data from four Scottish medical schools operating within a common regulatory framework. We evaluated longitudinal DA across undergraduate medical education in 1,512 medical students, exploring disability status, domicile, ethnicity, and gender. Here we report on the longitudinal effects of DA for these groups and the impact of DA on student rank.

Methods

Participants

Participants were undergraduate medical students who had attempted (but not necessarily passed) a major written (multiple choice question) assessment near the end of medical school. All institutions operated under the UK medical education system (2) and new graduates typically embarked on a two-year foundation training programme as a doctor.

In total, 1,512 medical students were eligible for inclusion in the study. To be eligible, a student had to (a) have attempted (but not necessarily passed) the final written assessment, (b) have made the attempt by the end of data collection and (c) have provided demographic information.

The 1,512 students represented 74% of all available participants within the period of this study. Excluded subjects were typically those who had exited medical school before final assessment, experienced an interruption of study, or those who had intercalated close to the end of the study period and so had not yet sat finals. Due to the complexity of discontinuation it is theoretically possible for a student to graduate up to nine years after starting a five-year programme, which makes confirmation of discontinuation challenging. Candidates who did not attempt final assessment prior to the end of the period of data collection are not included in any analyses presented here.

Table 1 summarises the partner schools, total sample sizes and assessments used. All schools offered five-year MBChBs with the option to intercalate and spend a year studying another subject in depth. Table 2 describes the participants according to important demographic characteristics. We report whether the candidate did or did not have a known disability, where they were domiciled before starting medical school, their ethnicity, and their gender. All recorded data was self-reported. For ethnicity and domicile, we aggregate data across many sub-categories into broad groups such as “international domicile” or “white.” While a more detailed breakdown would be helpful, the small numbers in many groups prohibit this. The demographic characteristics selected for study are based partly on the concept of a “protected characteristic” for which there is a legal obligation to promote equality within the UK (19), partly on demographic characteristics known to be important from past research, and partly on availability of data. To give two examples of data availability, marital status and sexual orientation had levels of missingness that were too high to achieve necessary levels of power. The four categories described here represent all those selected for full analysis and all analyses have sufficient power to detect medium effects.

[Insert Table 1 about here]

[Insert Table 2 about here]

Data Protection and ethics

This project represented a considerable challenge under data protection legislation and required a careful and thorough evaluation of ethical issues. To ensure data protection, a designated team member undertook an honorary contract with each partner and worked in tandem with a data custodian at that school. This meant individualised data was never transferred outside of the school servers, and a thorough anonymisation protocol was used to verify that no “unique” combinations could identify candidates from their data patterns. Ethical approval was granted by the ethics committee for the College of Medicine and Veterinary Medicine at the University of Edinburgh, and then separately approved by an ethics board and a data protection officer at each of the other schools. Prior to data analysis, all partners agreed to disseminate the results in public and to representatives of the study population: in this case, medical student organisations.

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Statistical analyses

Each medical school has a unique curriculum and assessment environment. We investigate written assessment as the most comparable form of assessment, as the available clinical examinations vary considerably across the schools in both timing and format. To allow like-for-like comparisons across different written assessments we converted each cohort of data to z-scores. (20)

A z-score is a standardised measurement, where a score of zero indicates the candidate has received exactly the mean mark on the assessment, and a score of +/ - 1 indicates they have received a mark one standard deviation above or below the mean respectively. This is analytically helpful because it allows for comparisons where relative (rather than absolute) differences are important. If a candidate from one medical school receives a mark of 75, and a candidate from another medical school receives a mark of 70 on two different assessments, it is difficult to know who is more capable. But if the z-score for each candidate is 0, this indicates they are of the same level of ability *relative to their peers* and that they are both average.

We used the Shapiro-Wilk test to model residual values to test for normality. (21) Although the normality parameters were violated ($W = 0.99$, $p < 0.001$) further investigation suggested that parametric testing would still be more appropriate as parametric tests are more effective at minimising the risk of false positives where the group sample sizes and standard deviations vary across groups. (22) Sample sizes were sufficient to detect small effects at 80% power for ethnicity, gender and domicile, whereas for disability status the unequal group sizes and small numbers of students self-reporting a disability allowed for only medium effects at 80% power. (23) Due to the low sample sizes *within* each medical school it was not feasible to compare inter-medical school variability with sufficient power. We used Welch's t-test for significance testing as a more robust alternative to other t-tests. (24) All analyses were carried out using *R*. (25)

Design choices

We made several design choices that influence the final dataset. Most importantly, by only including candidates who reach final assessment we exclude the majority of those who experienced major difficulties early in their studies. However, the only alternative is to either measure graduation rates, which prevents granular analyses as the overwhelming majority of students pass medical school (26), or attempt some form of imputation to estimate final performance of candidates who never reached that stage of education, with significant uncertainty over the accuracy of such estimates. We opt for a simple approach of reporting data only where fully available. One consequence of this is that variability is higher in final assessment than in first year, with more candidates performing poorly, so most z-score change values were negative. For example, it would be possible for a candidate to receive an A in first year and an F in final year and participate in our study, but it would not be possible for the reverse to be true – unless the student successfully resat assessment *and* then completed within the specified timeframe. This can be considered a form of "survival bias" and approaches to the problem always require trade-offs. (27)

Following this, we carried out a number of comparisons. Firstly, we calculated the z-score for each student in their first year and then final assessment. We explored the equivalence of school. We compared z-score *change* between groups to see whether attainment gaps were growing or shrinking during medical school. Finally, we ranked all candidates to see who would appear in either the top or bottom decile for the final assessment.

Results

We first tested whether the performance profiles of each school were sufficiently similar to pool data into a single sample. We compared the shapes of the distributions, frequencies of outliers, and the overall variability of each cohort. After confirming the equivalence of the cohorts, we pooled all data into a combined sample of 1,512 students.

Table 3 provides a summary of (a) the z-score for each demographic characteristic per assessment, (b) the relative change in z-score over time and (c) whether the z-score change *between* groups is significant. For the present study, we are not interested in the attainment gap at either the start or end of medical school – but whether the magnitude of the gap changes over time. We found that the gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). No significant differences were observed for candidates with vs. without a known disability.

[Insert Table 3 about here]

For the three significant analyses, non-white, internationally domiciled, and male candidates achieved a relatively higher score at the start of medical school. By the end of medical school, they were respectively behind white, Scottish-domiciled, and female students. The effect size was medium when testing ethnicity and domicile, and small for testing gender. In summary, non-white, internationally domiciled, and male students experienced a relative decline in their performance at medical school which cannot be explained by low attainment before or in the first year of medical school.

Finally, we estimated how often medical students of different demographics would appear in the top and bottom decile based on their z-scores vs. their expected frequencies based purely on how many existed in each category. Table 4 summarises the details.

[Insert Table 4 about here]

Decile 1 is the highest scoring decile and decile 10 is the lowest scoring decile. Students with a known disability, Scottish students and non-white students are over-represented in the bottom decile and under-represented in the top decile. Students with no known disability, and white students, are over-represented in the top decile and under-represented in the bottom decile. International students, and male students, are over-represented in *both* the top and bottom decile. Female students are under-represented in the top and bottom decile.

This analysis shows that many groups exhibit differences not just in mean performance, but also in variability, with some candidates being under- and over- represented at the extremes of the distribution. Demographic characteristics influence the likelihood of a candidate appearing in any given decile.

Discussion

Statement of principal findings

Differential Attainment exists within Scottish medical schools, with small to medium effects. The analysis described here demonstrates both the considerable difficulty in organising datasets to longitudinally investigate DA, and the ongoing importance of such work. Even among successful medical students – and the overwhelming majority of those described in the present dataset have become doctors – DA exists. The fact that many attainment gaps grow during medical school suggests educational factors within medical schools may promote DA.

Strengths and weaknesses of the study

It is important not to over-state the findings. Small to medium effect sizes are consequential and impact student education, but there remains considerable variance between students of all groups. In this dataset, candidates across the attainment continuum were present in every group. In addition, the core purpose of medical education – graduating a safe doctor – has been met for almost all participants in the dataset. The gaps observed here must be placed in this context. Finally, as until now we have operated in an environment with almost no published data, there is a risk that organisations which attempt to directly engage with the problem of DA are criticised for the differences they reveal, which may in turn drive reluctance to explore the issue in depth. It is important that stakeholders support the exploration of DA across the sector.

This study represents a novel attempt to understand DA not as a fixed factor, but as a changing influence on student performance and behaviour. The sample size and range suggest we can be confident the findings are potentially generalisable to other UK medical schools. By opting for a straightforward methodology, we believe the findings are robust and can inform future policy.

Despite this, there are limitations. The challenges of organising a longitudinal study using data from a range of institutions with varying outcome measures should not be understated. We have made design choices – such as excluding those who failed before reaching finals – which may influence the pattern of results. The lack of a shared, standardised assessment across schools required the use of z-scores (or an equivalent method) and the presence of a standardised assessment, such as the forthcoming UK Medical Licensing Assessment, would have greatly simplified the analysis. (28)

Data collection was challenging, and it was clear that there was no expectation during data creation that assessment-level data would be required five or ten years after the assessment was sat. Medical education data should be thought of as “perishable” – it is possible that even relatively recent data is being lost, overwritten, or rendered inaccessible. If medical educators wish to investigate DA across time it is critical that better data collection practices are implemented, and historic data sources should be secured and documented in national-level databases. (29) The alternative is that we may establish excellent prospective analyses for which we will have no useful data for up to a decade.

Comparison with other studies and unanswered questions

Our findings support and extend past work exploring DA in postgraduate medical education. (5,7,8,14) Importantly, our study also confirms that we remain unclear, as a sector, on the mechanisms behind DA. (11,12) All organisations involved in medical education must proactively consider how they approach fairness in medical education, and evaluate the impact of DA.

Besides the important future work of expanding such analyses to currently understudied groups and other institutions, we reiterate that the mechanisms of DA remain poorly understood. Medical educators must develop an awareness of the everyday factors that influence DA to mitigate them for students and trainees everywhere.

Implications and conclusions

This study demonstrates that DA is an important factor within undergraduate medical education, influencing students from a diverse range of groups. Significant additional work is required to better understand and deal with the practical challenges of DA in medical education.

References

1. Tsouroufli M, Malcolm I. Equality, diversity and fairness in medical education: international perspectives. Med Educ. 2015 Jan 1;49(1):4–6.

2. General Medical Council. Outcomes for Graduates. Manchester: General Medical Council; 2015.

3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: AERA; 2014.

4. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Medical Education. 2016;50(1):36–60.

5. MacKenzie RK, Cleland JA, Ayansina D, Nicholson S. Does the UKCAT predict performance on exit from medical school? A national cohort study. BMJ Open. 2016 Oct 1;6(10):e011313.

6. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ. 2011 Mar 8;342:d901.

7. Linton S. Taking the difference out of attainment. BMJ. 2020;368:m438.

8. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, et al. Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature. Journal of General Internal Medicine. 2019 May 1;34(5):712–9.

9. Woolf K. Differential attainment in medical education and training. BMJ. 2020;368:m339.

10. Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. BMC medicine. 2017;15(1):1–11.

11. The Queen on the application of Bapio Action Ltd [Cliamant] v Royal College of General Practitioners [First Defendant] and General Medical Council [Second Defendant], in the High Court of Justice, Queen's Bench Division, The Administrative Court. 10th April 2014. EWHC 1416 (Admin) 2014, Available at http://www.rcgp.org.uk/news/2014/may/~/media/Files/News/Judicial-Review-Judgment-14-April-2014.ashx.

12. Hope D, Adamson K, McManus IC, Chis L, Elder A. Using Differential Item Functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. BMC Medical Education. 2020;18(1):64.

13. Cleland J, Palma TF. "Aspirations of people who come from state education are different": how language reflects social exclusion in medical education. Advances in Health Sciences Education. 2018;23(3):513–31.

14. Woolf K, Rich A, Viney R, Needleman S, Griffin A. Perceived causes of differential attainment in UK postgraduate medical training: a national qualitative study. BMJ Open. 2016;6(11):e013429.

15. Atewologun D, Kline R, Ochieng M. Fair to refer? Reducing disproportionality in fitness to practise concerns reported to the GMC [Internet]. 2019. Available from: https://www.gmc-

uk.org/about/what-we-do-and-why/data-and-research/research-and-insight-archive/fair-to-refer

16. Kristoffersson E, Diderichsen S, Verdonk P, Lagro-Janssen T, Hamberg K, Andersson J. To select or be selected – gendered experiences in clinical training affect medical students' specialty preferences. BMC Medical Education. 2018;18(1):268.

17. McManus IC, Elder A, de Champlain A, Dacre J, Mollon J, Chis L. Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. BMC Medicine. 2008;6:5.

18. Devine OP, Harborne AC, McManus IC. Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. BMC medical education. 2015;15(1):146.

19. Davies C, Ferreira N, Morris A, Morris D. The Equality Act 2010: Five years on. International Journal of Discrimination and the Law. 2016;16(2–3):61–5.

20. Abdi H. Z-scores. In: Encyclopedia of measurement and statistics. Thousand Oaks (CA): Sage; 2007. p. 1055–8.

21. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611.

22. Zimmerman DW. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. The Journal of Experimental Education. 1998;67(1):55–68.

23. Cohen J. A Power Primer. Quantitative Methods in Psychology. 1992;112:155–9.

24. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology. 2017;30(1).

25. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996;5:299–314.

26. Arulampalam W, Naylor RA, Smith JP. A hazard model of the probability of medical school drop-out in the UK. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2004;167:157–78.

27. Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods. Am J Epidemiol. 2005;162(10):1016–23.

28. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Bere SR de. The medical licensing examination debate. Regulation & Governance. 2017;11(3):315–22.

29. Dowell J, Cleland J, Fitzpatrick S, McManus C, Nicholson S, Oppé T, et al. The UK medical education database (UKMED) what is it? Why and how might you use it? BMC medical education. 2018;18(1):1–8.

Table 1: Participants, data ranges and assessments used

School name	Sample size	Data range	First year assessment	Final assessment
University of Aberdeen	104	2014/2017	MCQ and SAQ	MCQ and SAQ
University of Dundee	202	2013/2016 & 2014/2017	MCQ	MCQ
University of Edinburgh	871	2009/2013, 2010/2014, 2011/2015, 2012/2016 & 2013/2017	MCQ and SAQ	MCQ
University of Glasgow	335	2014/2018 & 2015/2019	MCQ	MCQ

Note: Data range described the first/final year of assessment data for each cohort. "Multiple Choice Questions" (MCQs) requires students to select the correct answer from a series of options. "Short Answer Questions" (SAQs) require students to type or write a short answer. All assessments were written rather than clinical.

Table 2: Demographic characteristics of the study sample

Demographic Characteristic	Category	Institution	n		Total n
Disability	Known disability	Aberdeen	13	102	1512
		Dundee	13		
		Edinburgh	74		
		Glasgow	2		
	No known disability	Aberdeen	91	1410	
		Dundee	189		
		Edinburgh	797		
		Glasgow	333		
Domicile	EU (non-UK)	Aberdeen	2	44	1512
		Dundee	17		
		Edinburgh	14		
		Glasgow	11		
	International	Aberdeen	9	146	
		Dundee	12		
		Edinburgh	88		
		Glasgow	37		
	Rest of UK	Aberdeen	24	500	
		Dundee	40		
		Edinburgh	354		
		Glasgow	82		
	Scotland	Aberdeen	69	822	
		Dundee	133		
		Edinburgh	415		
		Glasgow	205		
Ethnicity	Non-white	Aberdeen	27	298	1512
		Dundee	21		
		Edinburgh	157		
		Glasgow	93		
	White	Aberdeen	77	1143	
		Dundee	165		
		Edinburgh	665		
		Glasgow	236		
	Unknown	Dundee	16	71	
		Edinburgh	49		
		Glasgow	6		
Gender	Female	Aberdeen	67	877	1512
		Dundee	129		
		Edinburgh	480		
		Glasgow	201		
	Male	Aberdeen	37	635	

Dundee	73
Edinburgh	391
Glasgow	134

Note: Candidates of "unknown" ethnicity, "EU (non-UK)" and "Rest of UK" domicile students are not included in any analyses described in the present study. All demographic characteristics relied on self-report data.

Table 3: Z-score change during medical school study

Demographic Characteristic	Category	First assessment (mean)	First assessment (SD)	Final assessment (mean)	Final assessment (SD)	Change (mean)	Change (SD)	Significance/CI
Disability	Known disability	-0.15	0.94	-0.38	0.73	-0.18	0.93	
	No known disability	0.09	0.89	-0.05	0.93	-0.1	0.95	
Domicile	International	0.46	0.83	-0.4	0.92	-0.57	0.92	* (-0.75 to -0.42)
	Scotland	-0.08	0.91	-0.05	0.9	0.01	0.97	
Ethnicity	Non-white	0.15	0.93	-0.34	1.06	-0.45	0.96	* (0.34 to 0.58)
	White	0.04	0.89	-0.02	0.88	0	0.92	
Gender	Female	0.03	0.89	-0.01	0.88	-0.03	0.93	* (0.08 to 0.27)
	Male	0.14	0.89	-0.2	0.98	-0.2	0.97	

Note: * indicates statistical significance at *p* = .001. 95% confidence intervals are given for significant results. For model values see text. Statistical significance indicates the relative attainment gap between categories changed significantly during the course of study.

Table 4: Rankings of top and bottom decile by demographic characteristic

Demographic characteristic	n	Category	n category	Percentage	Decile	n in decile	expected percentage	actual percentage
Disability	1512	Known disability	102	6.75	1	5	0.68	0.33
					10	14		0.93
		No known disability	1410	93.25	1	145	9.32	9.59
					10	136		8.99
Domicile	968	International	146	9.66	1	21	0.97	1.39
					10	19		1.26
		Scotland	822	54.37	1	78	5.44	5.16
					10	85		5.62
Ethnicity	1441	Non-white	298	19.71	1	24	1.97	1.59
					10	54		3.57
		White	1143	75.6	1	115	7.56	7.61
					10	92		6.08
Gender	1512	Female	877	58	1	85	5.8	5.62
					10	81		5.36
		Male	635	42	1	65	4.2	4.3
					10	69		4.56

Note: n indicates the total sample size for that characteristic, while n category indicates the sample size for the individual category. Percentage indicates the proportion of students from that category in the overall sample. Decile 1 is the highest (i.e. best performing) decile, decile 10 is the lowest (i.e. worst performing) decile. N in decile gives the number of candidates who actually appeared in that decile, and the difference between the expected and actual percentage shows whether the category is over- or under- represented.

BMJ Open

Measuring differential attainment: a longitudinal analysis of assessment results for 1,512 medical students at four Scottish medical schools

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046056.R1
Article Type:	Original research
Date Submitted by the Author:	18-Mar-2021
Complete List of Authors:	Hope, David; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit Dewar, Avril; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit Hothersall, Eleanor; University of Dundee, Medical Research Institute Leach , John ; Queen Elizabeth University Hospital, Institute of Neurological sciences Cameron, Isobel; University of Aberdeen, Applied Health Sciences (Mental Health) Jaap, Alan; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Ethics, Research methods
Keywords:	MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS, AUDIT, EDUCATION & TRAINING (see Medical Education & Training), HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™ Manuscripts

Measuring differential attainment: a longitudinal analysis of assessment results for 1,512 medical students at four Scottish medical schools

A retrospective analysis of change in attainment gaps during medical school study

David Hope[1], Avril Dewar[2], Eleanor J Hothersall[3], John Paul Leach[4], Isobel M Cameron[5] and Alan Jaap[6]

[1]Senior Lecturer in Medical Education, University of Edinburgh: david.hope@ed.ac.uk

[2]Fellow in Medical Education, University of Edinburgh: avril.dewar@ed.ac.uk

[3]Head of Undergraduate Medicine, University of Dundee: e.hothersall@dundee.ac.uk

[4]Head of Undergraduate Medicine, University of Glasgow, John.Leach@glasgow.ac.uk

[5]Senior Lecturer and Assessment Lead, University of Aberdeen, i.m.cameron@abdn.ac.uk

[6]Deputy Director of Teaching and Assessment Lead, University of Edinburgh, alan.jaap@ed.ac.uk

Address for Correspondence

Dr. David Hope

Medical Education Unit

The Chancellor's Building

The University of Edinburgh

College of Medicine and Veterinary Medicine

49 Little France Crescent, Edinburgh

EH16 4SB

Scotland, United Kingdom

david.hope@ed.ac.uk

Contributorship statement

Dr Cameron, Dr Hothersall, and Prof Leach were each responsible for sourcing data, describing the context, and exploring the results in their institutions. Avril Dewar was responsible sourcing data at her institution then collating all the data and running the initial analyses. Dr Hope organised the project, designed the analyses, was primarily responsible for writing the paper. Dr Hope is the guarantor for the content. Dr Jaap acted as supervisor for all the project work and reviewed the analyses.

All authors have separately reviewed the manuscript and provided input in developing the final analyses and paper.

Copyright

Competing interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors had financial support from the Scottish Medical Education Research Consortium (SMERC) for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Transparency statement

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned have been explained.

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Dissemination declaration

Results will be disseminated to the representatives of the study populations (medical student groups).

Role of the funding source

The Scottish Medical Education Research Consortium (SMERC) provided funding to allow the research project to take place. The funding was used to pay for administrator and researcher time to collate and analyse the data. The funder had no direct input into the analyses chosen or the reporting of the results. The researchers were independent from the funder, and all researchers had access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Ethical Approval

Ethical approval was granted by the ethics committee for the College of Medicine and Veterinary Medicine at the University of Edinburgh (reference: 2018/7), and then separately approved by an ethics board and a data protection officer at each of the other schools. All participants gave informed consent. Prior to data analysis, all partners agreed to disseminate the results in public and to representatives of the study population: in this case, medical student organisations. This information is reproduced in the main text.

Data sharing statement

Due to the sensitivity of the dataset – including confidential information on student demographics and assessment scores – we are unable to share raw data.

Word count

3,796

Abstract

Objective: To measure Differential Attainment (DA) among Scottish medical students and to explore whether attainment gaps increase or decrease during medical school.

Design: A retrospective analysis of undergraduate medical student performance on written assessment, measured at the start and end of medical school.

Setting: Four Scottish medical schools (Universities of Aberdeen, Dundee, Edinburgh, and Glasgow).

Participants: 1,512 medical students who attempted (but did not necessarily pass) final written assessment.

Main outcome measures: The study modelled the change in attainment gap during medical school for four student demographic categories (white/non-white, international/Scottish domiciled, male/female and with/without a known disability) to test whether the attainment gap grew, shrank, or remained stable during medical school. Separately, the study modelled the expected vs. actual frequency of different demographic groups in the top and bottom decile of the cohort.

Results: The attainment gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). International, non-white, and male students received higher marks than their comparison group at the start of medical school but lower marks by final assessment. No significant differences were observed for disability status. Students with a known disability, Scottish students and non-white students were over-represented in the bottom decile and under-represented in the top decile.

Conclusions: The tendency for attainment gaps to grow during undergraduate medical education suggests that educational factors at medical schools may – however inadvertently – contribute to Differential Attainment. It is of critical importance that medical schools investigate attainment gaps within their cohorts and explore potential underlying causes.

Article Summary

Strengths and limitations of this study

- This the largest study to date investigating longitudinal attainment gaps within undergraduate medical education
- By evaluating Differential Attainment longitudinally, the study tests whether attainment gaps are due to pre-existing differences or emerge during medical school
- The study has sufficient power to detect small/medium effects by pooling data from multiple cohorts and institutions
- All contributing schools were based in Scotland, and care should be taken when generalising to other contexts
- The study methodology cannot fully explain the mechanisms behind such attainment gaps

Introduction

Promoting fairness in assessment is a key priority. Success in medicine should be determined by ability rather than background characteristics like ethnicity, sex, or socioeconomic status. (1) There is an increasing emphasis on educational processes being "fair" to candidates of diverse backgrounds: besides the legal and regulatory requirements (2) there is growing acceptance that evaluating fairness should be a routine part of test construction and assessment. (3)

Despite this, candidates continue to experience different outcomes in medical education and training because they have characteristics which lead to them being treated differently by staff, students, and patients. The tendency for outcomes to vary in this fashion is usually termed Differential Attainment (DA). It influences every stage of medical education, and is a global phenomenon with similar problems manifesting in a range of contexts. (4,5) The varying treatment of some groups influences the likelihood of candidates completing medical school and affects selection methods. (6–8) Performance on measures of success at or just beyond graduation show a similar pattern (9,10) and, for example, ethnically white UK graduates are given higher marks than non-white UK graduates in postgraduate examinations with typically moderate (d = 0.22) effects. (11) After graduation, ethnically non-white and female doctors experience barriers to success on a range of professional and educational outcomes. (12–14) Students from underrepresented backgrounds are substantially less likely to be awarded high ratings from their clerkship directors, less likely to be given honours, and less likely to be given honour society membership. (15)

Such compelling evidence has led to calls to establish the mechanisms of DA, but this is challenging. Many historical assumptions – such as the idea that examiners are biased against some candidate groups – remain commonly cited despite evidence to the contrary. (16,17) Examiner bias does not appear to explain DA in postgraduate clinical examinations (18) or written assessment. (19) Qualitative research has emphasised a range of possible factors that can contribute to DA, including trust between trainers and trainees and the process by which those in difficulty are identified and referred to support networks. (20–22) Other research has suggested that unconscious biases may alter training pathways or assessment in the workplace. (4,13,23,24) Some authors now recommend a programmatic approach whereby each component of training is separately reviewed. (25)

As a result, evidence for the existence of DA is very strong but we have so far only a limited understanding of the mechanisms by which it operates or even whether DA increases or decreases with time spent in medical education. Compounding this, while a great deal of research has been carried out on access to medical school and postgraduate assessment, relatively little work has evaluated DA on assessment *during* medical school. In a large meta-analysis eleven of fourteen published studies examining undergraduate medical education used a single site and two of the remaining studies used only two sites. (11) Combined with the tendency to monitor attainment at only a single time point (typically finals) we know little of whether DA is of similar magnitude for different medical schools or remains stable during medical school.

This is an obvious limitation given the role of medical schools in providing the foundation of medical education and training. Due to the diversity of intakes, assessment choices, curriculum design and performance on postgraduate assessment (26,27) investigating DA at medical schools may help in several ways. By comparing different institutions, the effect of different recruitment strategies, curriculum types and policies on fairness in medical education can be explored. If the magnitude of DA is highly variable across institutions, it argues for a relatively larger role in medical school policy in creating DA. If DA remains consistent despite varying institutional contexts, it argues DA is either explained by factors outside of medical school control or that no current approaches are identifiably superior or inferior. By examining the data longitudinally, it becomes possible to explore whether DA increases or decreases over time. If DA is present from the earliest part of medical education, this suggests different mechanisms than if DA is minimally present at the beginning but then grows with time. Such work can therefore significantly improve medical education and support a fairer experience for doctors.

In this study we used data from four Scottish medical schools operating within a common regulatory framework. Our aim was to evaluate longitudinal DA across undergraduate medical education in 1,512 medical students, exploring disability status, domicile, ethnicity, and gender. Here we report on the longitudinal effects of DA for these groups and the impact of DA on student rank.

Methods

Participants

Participants were undergraduate medical students who had attempted (but not necessarily passed) a major written (multiple choice question) assessment near the end of medical school. All institutions operated under the UK medical education system (2) and new graduates typically embarked on a two-year foundation training programme as a doctor.

In total, 1,512 medical students were eligible for inclusion in the study. To be eligible, a student had to (a) have attempted (but not necessarily passed) the final written assessment, (b) have made the attempt by the end of data collection and (c) have provided demographic information.

The 1,512 students represented 74% of all available participants within the period of this study. Excluded subjects were typically those who had exited medical school before final assessment, experienced an interruption of study, or those who had intercalated close to the end of the study period and so had not yet sat finals. Due to the complexity of discontinuation it is theoretically possible for a student to graduate up to nine years after starting a five-year programme, which makes confirmation of discontinuation challenging. Candidates who did not attempt final assessment prior to the end of the period of data collection are not included in any analyses presented here.

Table 1 summarises the partner schools, total sample sizes and assessments used. All schools offered five-year MBChBs. The first two years of each programme involved an introduction to the fundamentals of medicine, anatomy, social issues around healthcare, and working with peers. Each programme offered an opportunity to intercalate, whereby candidates spent an additional year studying a topic in greater depth before returning to the core programme. In the later years, candidates rotated through a series of clinical placements to develop the skills and knowledge necessary to work as a junior doctor.

In each school, candidates sat a written assessment at the end of their first year. These featured Multiple Choice Questions (MCQs) and for two schools, Short Answer Questions (SAQs). For each question candidates were presented with a scenario and question. For MCQs candidates selected the correct answer from a list, whereas for SAQs candidates provided a short, written answer. The assessment was blueprinted based on programme learning outcomes and standard set by experts familiar with the curriculum.

Near the end of medical school, candidates sat another written assessment. Three schools delivered this in final year, while one (the University of Aberdeen) delivered it at the very end of the prefinal year. The blueprinting and standard setting process was the same as in the early assessment.

In each case, the assessments acted as a progression barrier: candidates needed to achieve a satisfactory mark to progress to either second year or graduation. A review by the authors identified that although there were some variations in curricula and teaching methods there were no significant differences in content and structure of assessments between programmes that would impact cross-school comparisons of DA.

Table 2 describes the participants according to important demographic characteristics. We report whether the candidate did or did not have a known disability, where they were domiciled before starting medical school, their ethnicity, and their gender. All recorded data was self-reported. For ethnicity and domicile, we aggregate data across many sub-categories into broad groups such as “Scottish domicile” or “white.” While a more detailed breakdown would be helpful, the small numbers in many groups prohibit this. The demographic characteristics selected for study are based partly on the concept of a “protected characteristic” for which there is a legal obligation to promote equality within the UK (28), partly on demographic characteristics known to be important from past research, and partly on availability of data. To give two examples of data availability, marital status and sexual orientation had levels of missingness that were too high to achieve necessary levels of power. The four categories described here (known/no known disability, international/Scottish domicile, non-white/white and female/male) represent all those selected for full analysis and all analyses have sufficient power to detect medium effects. We selected Scottish (as opposed to whole UK) domicile due to Scottish domiciled candidates having already experienced the Scottish legislative and educational framework and having selected a medical school relatively close to home. Furthermore, differences in the funding approach in Scotland compared to the rest of the UK made merging the two groups less defensible.

Socioeconomic status (SES) was recorded in the dataset in two forms. Firstly, candidates had the opportunity to list parental occupation. Over 90% of candidates did not fill this in. A second proxy for SES was candidate postcode, which can be converted into an index of multiple deprivation. (29) However, it was not possible to effectively compare Scottish, non-Scottish UK, and international measures of SES within a single dataset. As such we did not explore this covariate further in the present study.

[Insert Table 1 about here]

[Insert Table 2 about here]

Data Protection and ethics

This project represented a considerable challenge under data protection legislation and required a careful and thorough evaluation of ethical issues. To ensure data protection, a designated team member undertook an honorary contract with each partner and worked in tandem with a data custodian at that school. This meant individualised data was never transferred outside of the school servers, and a thorough anonymisation protocol was used to verify that no “unique” combinations could identify candidates from their data patterns. Ethical approval was granted by the ethics committee for the College of Medicine and Veterinary Medicine at the University of Edinburgh (reference: 2018/7), and then separately approved by an ethics board and a data protection officer at each of the other schools. All participants gave informed consent. Prior to data analysis, all partners agreed to disseminate the results in public and to representatives of the study population: in this case, medical student organisations.

When describing inequities researchers must ensure individuals are described fairly and appropriately, without discriminatory language. Throughout this paper, we have used language which shows that group membership itself does not *cause* an attainment gap and is never a direct determinant of performance, and instead likely reflects systemic societal issues. We have provided some additional references which may be helpful in exploring language choice when describing historically under-represented groups. (4,20)

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Statistical analyses

Each medical school has a locally designed curriculum and assessment environment. We investigate written assessment as the most comparable form of assessment, as the available clinical examinations vary considerably across the schools in both timing and format. To allow like-for-like comparisons across different written assessments we converted each cohort of data to z-scores. (30)

A z-score is a standardised measurement, where a score of zero indicates the candidate has received exactly the mean mark on the assessment, and a score of +/ - 1 indicates they have received a mark one standard deviation above or below the mean respectively. This is analytically helpful because it allows for comparisons where relative (rather than absolute) differences are important. If a candidate from one medical school receives a mark of 75, and a candidate from another medical school receives a mark of 70 on two different assessments, it is difficult to know who is more capable. But if the z-score for each candidate is 0, this indicates they are of the same level of ability *relative to their peers* and that they are both average.

We used the Shapiro-Wilk test to model residual values to test for normality. (31) Although the normality parameters were violated ($W = 0.99$, $p < 0.001$) further investigation suggested that parametric testing would still be more appropriate as parametric tests are more effective at minimising the risk of false positives where the group sample sizes and standard deviations vary across groups. (32) Sample sizes were sufficient to detect small effects at 80% power for ethnicity, gender and domicile, whereas for disability status the unequal group sizes and small numbers of students self-reporting a disability allowed for only medium effects at 80% power. (33) Due to the low sample sizes *within* each medical school it was not feasible to compare inter-medical school variability with sufficient power. Likewise, it was not possible to compare intersectional DA (e.g., ethnicity *and* gender). We used Welch's t-test for significance testing as a more robust alternative to other t-tests. (34) All analyses were carried out using *R*. (35)

Design choices

We made several design choices that influence the final dataset. Most importantly, by only including candidates who reach final assessment we exclude the majority of those who experienced major difficulties early in their studies. However, the only alternative is to either measure graduation rates, which prevents granular analyses as the overwhelming majority of students pass medical school (36), or attempt some form of imputation to estimate final performance of candidates who never reached that stage of education, with significant uncertainty over the accuracy of such estimates. We opt for a simple approach of reporting data only where fully available. One consequence of this is that variability is higher in final assessment than in first year, with more candidates performing poorly, so most z-score change values were negative. For example, it would be possible for a candidate to receive an A in first year and an F in final year and participate in our study, but it would not be possible for the reverse to be true – unless the student successfully resat assessment *and* then completed within the specified timeframe. This can be considered a form of "survival bias" and approaches to the problem always require trade-offs. (37)

To investigate survival bias, we compared the ratios of those who did versus those who did not provide final year assessment results for each group. For example, we compared the ratio of non-white/white completers against non-white/white non-completers. No differences in the ratios were detected for any studied group. This likely reflects the fact that non-completion (by the end of the present study) was due to a variety of factors and did not in itself indicate academic difficulty.

Following this, we carried out a number of comparisons. Firstly, we calculated the z-score for each student in their first year and then final assessment. We explored the equivalence of school. We compared z-score *change* between groups to see whether attainment gaps were growing or shrinking during medical school. Finally, we ranked all candidates to see who would appear in either the top or bottom decile for the final assessment.

Results

We first tested whether the performance profiles of each school were sufficiently similar to pool data into a single sample. We compared the shapes of the distributions, frequencies of outliers, and the overall variability of each cohort. After confirming the equivalence of the cohorts, we pooled all data into a combined sample of 1,512 students.

Table 3 provides a summary of (a) the z-score for each demographic characteristic per assessment, (b) the relative change in z-score over time and (c) whether the z-score change *between* groups is significant. For the present study, we are not interested in the attainment gap at either the start or end of medical school – but whether the magnitude of the gap changes over time. We found that the gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). No significant differences were observed for candidates with vs. without a known disability.

[Insert Table 3 about here]

For the three significant analyses, non-white, internationally domiciled, and male candidates were awarded a relatively higher score at the start of medical school. By the end of medical school, they were respectively awarded a lower score then white, Scottish-domiciled, and female students. The effect size was medium when testing ethnicity and domicile, and small for testing gender. In summary, non-white, internationally domiciled, and male students experienced a relative decline in their achieved marks at medical school which cannot be explained by low attainment before or in the first year of medical school.

Finally, we estimated how often medical students of different demographics would appear in the top and bottom decile based on their z-scores vs. their expected frequencies based purely on how many existed in each category. Table 4 summarises the details.

[Insert Table 4 about here]

Decile 1 is the highest scoring decile and decile 10 is the lowest scoring decile. Students with a known disability, Scottish students and non-white students are over-represented in the bottom decile and under-represented in the top decile. Students with no known disability, and white students, are over-represented in the top decile and under-represented in the bottom decile. International students, and male students, are over-represented in *both* the top and bottom decile. Female students are under-represented in the top and bottom decile.

This analysis shows that many groups exhibit differences not just in mean performance, but also in variability, with some candidates being under- and over- represented at the extremes of the distribution.

Discussion

Statement of principal findings

Differential Attainment exists within Scottish medical schools, with small to medium effects. The analysis described here demonstrates both the considerable difficulty in organising datasets to longitudinally investigate DA, and the ongoing importance of such work. Even among successful medical students – and the overwhelming majority of those described in the present dataset have become doctors – DA exists. The fact that many attainment gaps grow during medical school suggests educational factors within medical schools may promote DA.

Strengths and weaknesses of the study

It is important not to over-state the findings. Small to medium effect sizes are consequential and impact student education, but there remains considerable variance between students of all groups. In this dataset, candidates across the attainment continuum were present in every group. In addition, the core purpose of medical education – graduating a safe doctor – has been met for almost all participants in the dataset. The gaps observed here must be placed in this context. Finally, as until now we have operated in an environment with almost no published data, there is a risk that organisations which attempt to directly engage with the problem of DA are criticised for the differences they reveal, which may in turn drive reluctance to explore the issue in depth. It is important that stakeholders support the exploration of DA across the sector.

This study represents a novel attempt to understand DA not as a fixed factor, but as a changing influence on student performance and behaviour. The sample size and range suggest we can be confident the findings are potentially generalisable to other UK medical schools. By opting for a straightforward methodology, we believe the findings are robust and can inform future policy.

Despite this, there are limitations. The challenges of organising a longitudinal study using data from a range of institutions with varying outcome measures should not be understated. We have made design choices – such as excluding those who failed before reaching finals – which may influence the pattern of results. Due to the relatively small sample sizes of some groups, it was not possible to explore "intersectional" DA for e.g. candidates who were non-white and female. (38) Due to the nature of the available data on SES, we were not able to include SES as a covariate in the present study. All candidate demographics were self-reported, and so some information could theoretically be inaccurate. While we consider the curricula and assessment of the institutions to be sufficiently similar to allow for a combined analysis, it is possible that local factors may have created some unidentified sources of variance.

The lack of a shared, standardised assessment across schools required the use of z-scores (or an equivalent method) and the presence of a standardised assessment, such as the forthcoming UK Medical Licensing Assessment, would have greatly simplified the analysis. (39)

Data collection was challenging, and it was clear that there was no expectation during data creation that assessment-level data would be required five or ten years after the assessment was sat. Medical education data should be thought of as "perishable" – it is possible that even relatively recent data is being lost, overwritten, or rendered inaccessible. If medical educators wish to investigate DA across time it is critical that better data collection practices are implemented, and historic data sources should be secured and documented in national-level databases. (40) The alternative is that we may establish excellent prospective analyses for which we will have no useful data for up to a decade.

Comparison with other studies and unanswered questions

DA exists across medical education systems across the world and should always be considered when designing teaching and assessment. (4,5) Our findings support and extend past work exploring DA in postgraduate medical education. (9,12,13,21) and at medical school. (15,24) Importantly, our study also confirms that we remain unclear, as a sector, on the mechanisms behind DA. (18,19) All organisations involved in medical education must proactively consider how they approach fairness in medical education, and evaluate the impact of DA.

The limitations described above are logical opportunities for future work. Exploring the impact of SES, analysing intersectional characteristics, and studying those who do not graduate may offer insights into both the scope and mechanisms of DA. Exploring candidate domicile in a more granular fashion (such as measuring the distance between home and their selected medical school) may be helpful, especially alongside measurements of SES. Importantly, the design challenges highlighted here will persist until institutions develop rigorous frameworks to investigate long-term changes in student performance.

Implications and conclusions

The present study demonstrates DA changes in magnitude during undergraduate medical education. Combined with evidence that candidates of some groups are less likely to be given awards (15) and more likely to experience prejudice (24), it is very plausible that some of the mechanisms of DA are located in, or caused by, aspects of medical education within medical schools. As such, institutions must consider the possibility that their actions contribute to DA and develop appropriate policies for investigation and correction. (14)

References

1. Tsouroufli M, Malcolm I. Equality, diversity and fairness in medical education: international perspectives. Med Educ. 2015 Jan 1;49(1):4–6.

2. General Medical Council. Outcomes for Graduates. Manchester: General Medical Council; 2015.

3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: AERA; 2014.

4. Murphy M. Teaching and Learning About Sexual Diversity Within Medical Education: the Promises and Pitfalls of the Informal Curriculum. Sex Res Soc Policy. 2019 Mar;16(1):84–99.

5. Prideaux D, Roberts C, Eva K, Centeno A, Mccrorie P, Mcmanus C, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011;33(3):215–23.

6. O'Neill L, Hartvigsen J, Wallstedt B, Korsholm L, Eika B. Medical school dropout - testing at admission versus selection by highest grades as predictors. Medical Education. 2011;45:1111–20.

7. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Medical Education. 2016;50(1):36–60.

8. Cliffordson C. Selection Effects on Applications and Admissions to Medical Education with Regular and Step-Wise Admission Procedures. Scandinavian Journal of Educational Research. 2006 Sep 1;50(4):463–82.

9. MacKenzie RK, Cleland JA, Ayansina D, Nicholson S. Does the UKCAT predict performance on exit from medical school? A national cohort study. BMJ Open. 2016 Oct 1;6(10):e011313.

10. Pershing S, Co JPT, Katznelson L. The New USMLE Step 1 Paradigm: An Opportunity to Cultivate Diversity of Excellence. Academic Medicine. 2020 Sep 1;95(9):1325–8.

11. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ. 2011 Mar 8;342:d901.

12. Linton S. Taking the difference out of attainment. BMJ. 2020;368:m438.

13. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, et al. Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature. Journal of General Internal Medicine. 2019 May 1;34(5):712–9.

14. Ufomata E, Merriam S, Puri A, Lupton K, LeFrancois D, Jones D, et al. A Policy Statement of the Society of General Internal Medicine on Tackling Racism in Medical Education: Reflections on the Past and a Call to Action for the Future. J GEN INTERN MED [Internet]. 2021 Jan 22 [cited 2021 Mar 8]; Available from: https://doi.org/10.1007/s11606-020-06445-2

15. Teherani A, Hauer KE, Fernandez A, King TEJ, Lucey C. How Small Differences in Assessed Clinical Performance Amplify to Large Differences in Grades and Awards: A Cascade With

Serious Consequences for Students Underrepresented in Medicine. Academic Medicine. 2018 Sep;93(9):1286–92.

16. Woolf K. Differential attainment in medical education and training. BMJ. 2020;368:m339.

17. Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. BMC medicine. 2017;15(1):1–11.

18. The Queen on the application of Bapio Action Ltd [Cliamant] v Royal College of General Practitioners [First Defendant] and General Medical Council [Second Defendant], in the High Court of Justice, Queen's Bench Division, The Administrative Court. 10th April 2014. EWHC 1416 (Admin) 2014, Available at http://www.rcgp.org.uk/news/2014/may/~/media/Files/News/Judicial-Review-Judgment-14-April-2014.ashx.

19. Hope D, Adamson K, McManus IC, Chis L, Elder AT. Using Differential Item Functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. BMC Medical Education. 2018;18(1):64.

20. Cleland J, Palma TF. "Aspirations of people who come from state education are different": how language reflects social exclusion in medical education. Advances in Health Sciences Education. 2018;23(3):513–31.

21. Woolf K, Rich A, Viney R, Needleman S, Griffin A. Perceived causes of differential attainment in UK postgraduate medical training: a national qualitative study. BMJ Open. 2016;6(11):e013429.

22. Atewologun D, Kline R, Ochieng M. Fair to refer? Reducing disproportionality in fitness to practise concerns reported to the GMC [Internet]. 2019. Available from: https://www.gmc-uk.org/about/what-we-do-and-why/data-and-research/research-and-insight-archive/fair-to-refer

23. Kristoffersson E, Diderichsen S, Verdonk P, Lagro-Janssen T, Hamberg K, Andersson J. To select or be selected – gendered experiences in clinical training affect medical students' specialty preferences. BMC Medical Education. 2018;18(1):268.

24. Cheng L-F, Yang H-C. Learning about gender on campus: an analysis of the hidden curriculum for medical students. Medical Education. 2015;49(3):321–31.

25. Karani R, Varpio L, May W, Horsley T, Chenault J, Miller KH, et al. Commentary: Racism and Bias in Health Professions Education: How Educators, Faculty Developers, and Researchers Can Make a Difference. Academic Medicine. 2017 Nov;92(11S):S1.

26. McManus IC, Elder A, de Champlain A, Dacre J, Mollon J, Chis L. Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. BMC Medicine. 2008;6:5.

27. Devine OP, Harborne AC, McManus IC. Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. BMC medical education. 2015;15(1):146.

28. Davies C, Ferreira N, Morris A, Morris D. The Equality Act 2010: Five years on. International Journal of Discrimination and the Law. 2016;16(2–3):61–5.

29. Noble M, Wright G, Smith G, Dibben C. Measuring Multiple Deprivation at the Small-Area Level. Environ Plan A. 2006 Jan 1;38(1):169–85.

30. Abdi H. Z-scores. In: Encyclopedia of measurement and statistics. Thousand Oaks (CA): Sage; 2007. p. 1055–8.

31. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611.

32. Zimmerman DW. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. The Journal of Experimental Education. 1998;67(1):55–68.

33. Cohen J. A Power Primer. Quantitative Methods in Psychology. 1992;112:155–9.

34. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch’s t-test instead of Student’s t-test. International Review of Social Psychology. 2017;30(1).

35. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996;5:299–314.

36. Arulampalam W, Naylor RA, Smith JP. A hazard model of the probability of medical school drop-out in the UK. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2004;167:157–78.

37. Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods. Am J Epidemiol. 2005;162(10):1016–23.

38. Morrison N, Chimkupete P. Double jeopardy: black and female in medicine. The Clinical Teacher. 2020;17(5):566–8.

39. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Bere SR de. The medical licensing examination debate. Regulation & Governance. 2017;11(3):315–22.

40. Dowell J, Cleland J, Fitzpatrick S, McManus C, Nicholson S, Oppé T, et al. The UK medical education database (UKMED) what is it? Why and how might you use it? BMC medical education. 2018;18(1):1–8.

505 **Table 1: Participants, data ranges and assessments used**

506

School name	Sample size	Data range	First year assessment	Final assessment
University of Aberdeen	104	2014/2017	MCQ and SAQ	MCQ and SAQ
University of Dundee	202	2013/2016 & 2014/2017	MCQ	MCQ
University of Edinburgh	871	2009/2013, 2010/2014, 2011/2015, 2012/2016 & 2013/2017	MCQ and SAQ	MCQ
University of Glasgow	335	2014/2018 & 2015/2019	MCQ	MCQ

507

508 Note: Data range described the first/final year of assessment data for each cohort. "Multiple Choice
509 Questions" (MCQs) requires students to select the correct answer from a series of options. "Short
510 Answer Questions" (SAQs) require students to type or write a short answer. All assessments were
511 written rather than clinical.

512

Table 2: Demographic characteristics of the study sample

Demographic Characteristic	Category	Institution	n	Total n	
Disability	Known disability	Aberdeen	13	102	1512
		Dundee	13		
		Edinburgh	74		
		Glasgow	2		
	No known disability	Aberdeen	91	1410	
		Dundee	189		
		Edinburgh	797		
		Glasgow	333		
Domicile	EU (non-UK)	Aberdeen	2	44	1512
		Dundee	17		
		Edinburgh	14		
		Glasgow	11		
	International	Aberdeen	9	146	
		Dundee	12		
		Edinburgh	88		
		Glasgow	37		
	Rest of UK	Aberdeen	24	500	
		Dundee	40		
		Edinburgh	354		
		Glasgow	82		
	Scotland	Aberdeen	69	822	
		Dundee	133		
		Edinburgh	415		
		Glasgow	205		
Ethnicity	Non-white	Aberdeen	27	298	1512
		Dundee	21		
		Edinburgh	157		
		Glasgow	93		
	White	Aberdeen	77	1143	
		Dundee	165		
		Edinburgh	665		
		Glasgow	236		
	Unknown	Dundee	16	71	
		Edinburgh	49		
		Glasgow	6		
Gender	Female	Aberdeen	67	877	1512
		Dundee	129		
		Edinburgh	480		
		Glasgow	201		
	Male	Aberdeen	37	635	

Dundee	73
Edinburgh	391
Glasgow	134

Note: Candidates of "unknown" ethnicity, "EU (non-UK)" and "Rest of UK" domicile students are not included in any analyses described in the present study. All demographic characteristics relied on self-report data.

Table 3: Z-score change during medical school study

Demographic Characteristic	Category	First assessment (mean)	First assessment (SD)	Final assessment (mean)	Final assessment (SD)	Change (mean)	Change (SD)	Significance/CI
Disability	Known disability	-0.15	0.94	-0.38	0.73	-0.18	0.93	
	No known disability	0.09	0.89	-0.05	0.93	0.1	0.95	
Domicile	International	0.46	0.83	-0.4	0.92	-0.57	0.92	* (-0.75 to -0.42)
	Scotland	-0.08	0.91	-0.05	0.9	0.01	0.97	
Ethnicity	Non-white	0.15	0.93	-0.34	1.06	-0.45	0.96	* (0.34 to 0.58)
	White	0.04	0.89	-0.02	0.88	0	0.92	
Gender	Female	0.03	0.89	-0.01	0.88	-0.03	0.93	* (0.08 to 0.27)
	Male	0.14	0.89	-0.2	0.98	0.2	0.97	

Note: * indicates statistical significance at *p* = .001. 95% confidence intervals are given for significant results. For model values see text. Statistical significance indicates the relative attainment gap between categories changed significantly during the course of study.

Table 4: Rankings of top and bottom decile by demographic characteristic

Demographic characteristic	n	Category	n category	Percentage	Decile	n in decile	expected percentage	actual percentage
Disability	1512	Known disability	102	6.75	1	5	0.68	0.33
					10	14		0.93
		No known disability	1410	93.25	1	145	9.32	9.59
					10	136		8.99
Domicile	968	International	146	9.66	1	21	0.97	1.39
					10	19		1.26
		Scotland	822	54.37	1	78	5.44	5.16
					10	85		5.62
Ethnicity	1441	Non-white	298	19.71	1	24	1.97	1.59
					10	54		3.57
		White	1143	75.6	1	115	7.56	7.61
					10	92		6.08
Gender	1512	Female	877	58	1	85	5.8	5.62
					10	81		5.36
		Male	635	42	1	65	4.2	4.3
					10	69		4.56

Note: n indicates the total sample size for that characteristic, while n category indicates the sample size for the individual category. Percentage indicates the proportion of students from that category in the overall sample. Decile 1 is the highest (i.e. best performing) decile, decile 10 is the lowest (i.e. worst performing) decile. N in decile gives the number of candidates who actually appeared in that decile, and the difference between the expected and actual percentage shows whether the category is over- or under- represented.

institutions must consider the possibility that their actions contribute to DA and develop appropriate policies for investigation and correction**STROBE 2007 (v4) checklist of items to be included in reports of observational studies in epidemiology***

Checklist for cohort, case-control, and cross-sectional studies (combined)

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1/1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	5/70
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	7/109
Objectives	3	State specific objectives, including any pre-specified hypotheses	8/158
Methods			
Study design	4	Present key elements of study design early in the paper	8/164
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8/164 – 11/292
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	8/164
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	9/200
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	11/246 – 12 /268
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	9/179
Bias	9	Describe any efforts to address potential sources of bias	9/195 & 10/226
Study size	10	Explain how the study size was arrived at	8/168
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9/203 & 11/246
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	11/246
		(*b*) Describe any methods used to examine subgroups and interactions	12/270
		(*c*) Explain how missing data were addressed	12/283
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed	12/283

		Case-control study—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	
		(*e*) Describe any sensitivity analyses	NA
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8/164 & 12/294
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	9/200
		(b) Indicate number of participants with missing data for each variable of interest	9/200
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	9/179
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	NA
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	9/200
		Cross-sectional study—Report numbers of outcome events or summary measures	NA
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	NA
		(*b*) Report category boundaries when continuous variables were categorized	NA
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	12/294
Discussion			
Key results	18	Summarise key results with reference to study objectives	14/330
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	14/350
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	15/385
Generalisability	21	Discuss the generalisability (external validity) of the study results	15/371
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	3/57

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

Measuring differential attainment: a longitudinal analysis of assessment results for 1,512 medical students at four Scottish medical schools

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046056.R2
Article Type:	Original research
Date Submitted by the Author:	19-Jul-2021
Complete List of Authors:	Hope, David; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit Dewar, Avril; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit Hothersall, Eleanor; University of Dundee, Medical Research Institute Leach , John ; Queen Elizabeth University Hospital, Institute of Neurological sciences Cameron, Isobel; University of Aberdeen, Applied Health Sciences (Mental Health) Jaap, Alan; The University of Edinburgh College of Medicine and Veterinary Medicine, Medical Education Unit
Primary Subject Heading:	Medical education and training
Secondary Subject Heading:	Ethics, Research methods
Keywords:	MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS, AUDIT, EDUCATION & TRAINING (see Medical Education & Training), HEALTH SERVICES ADMINISTRATION & MANAGEMENT

Measuring differential attainment: a longitudinal analysis of assessment results for 1,512 medical students at four Scottish medical schools

A retrospective analysis of change in attainment gaps during medical school study

David Hope[1], Avril Dewar[2], Eleanor J Hothersall[3], John Paul Leach[4], Isobel M Cameron[5] and Alan Jaap[6]

[1]Senior Lecturer in Medical Education, University of Edinburgh: david.hope@ed.ac.uk

[2]Fellow in Medical Education, University of Edinburgh: avril.dewar@ed.ac.uk

[3]Head of Undergraduate Medicine, University of Dundee: e.hothersall@dundee.ac.uk

[4]Head of Undergraduate Medicine, University of Glasgow, John.Leach@glasgow.ac.uk

[5]Senior Lecturer and Assessment Lead, University of Aberdeen, i.m.cameron@abdn.ac.uk

[6]Deputy Director of Teaching and Assessment Lead, University of Edinburgh, alan.jaap@ed.ac.uk

Address for Correspondence

Dr. David Hope

Medical Education Unit

The Chancellor's Building

The University of Edinburgh

College of Medicine and Veterinary Medicine

49 Little France Crescent, Edinburgh

EH16 4SB

Scotland, United Kingdom

david.hope@ed.ac.uk

Contributorship statement

Dr Cameron, Dr Hothersall, and Prof Leach were each responsible for sourcing data, describing the context, and exploring the results in their institutions. Avril Dewar was responsible sourcing data at her institution then collating all the data and running the initial analyses. Dr Hope organised the project, designed the analyses, was primarily responsible for writing the paper. Dr Hope is the guarantor for the content. Dr Jaap acted as supervisor for all the project work and reviewed the analyses.

All authors have separately reviewed the manuscript and provided input in developing the final analyses and paper.

Copyright

Competing interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors had financial support from the Scottish Medical Education Research Consortium (SMERC) for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Transparency statement

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned have been explained.

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Dissemination declaration

Results will be disseminated to the representatives of the study populations (medical student groups).

Role of the funding source

The Scottish Medical Education Research Consortium (SMERC) provided funding to allow the research project to take place. The funding was used to pay for administrator and researcher time to collate and analyse the data. The funder had no direct input into the analyses chosen or the reporting of the results. The researchers were independent from the funder, and all researchers had access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Ethical Approval

Ethical approval was granted by the ethics committee for the College of Medicine and Veterinary Medicine at the University of Edinburgh (reference: 2018/7), and then separately approved by an ethics board and a data protection officer at each of the other schools. All participants gave informed consent. Prior to data analysis, all partners agreed to disseminate the results in public and to representatives of the study population: in this case, medical student organisations. This information is reproduced in the main text.

Data sharing statement

Due to the sensitivity of the dataset – including confidential information on student demographics and assessment scores – we are unable to share raw data.

Word count

3,796

Abstract

Objective: To measure Differential Attainment (DA) among Scottish medical students and to explore whether attainment gaps increase or decrease during medical school.

Design: A retrospective analysis of undergraduate medical student performance on written assessment, measured at the start and end of medical school.

Setting: Four Scottish medical schools (Universities of Aberdeen, Dundee, Edinburgh, and Glasgow).

Participants: 1,512 medical students who attempted (but did not necessarily pass) final written assessment.

Main outcome measures: The study modelled the change in attainment gap during medical school for four student demographic categories (white/non-white, international/Scottish domiciled, male/female and with/without a known disability) to test whether the attainment gap grew, shrank, or remained stable during medical school. Separately, the study modelled the expected vs. actual frequency of different demographic groups in the top and bottom decile of the cohort.

Results: The attainment gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). International, non-white, and male students received higher marks than their comparison group at the start of medical school but lower marks by final assessment. No significant differences were observed for disability status. Students with a known disability, Scottish students and non-white students were over-represented in the bottom decile and under-represented in the top decile.

Conclusions: The tendency for attainment gaps to grow during undergraduate medical education suggests that educational factors at medical schools may – however inadvertently – contribute to Differential Attainment. It is of critical importance that medical schools investigate attainment gaps within their cohorts and explore potential underlying causes.

Article Summary

Strengths and limitations of this study

- This the largest study to date investigating longitudinal attainment gaps within undergraduate medical education
- By evaluating Differential Attainment longitudinally, the study tests whether attainment gaps are due to pre-existing differences or emerge during medical school
- The study has sufficient power to detect small/medium effects by pooling data from multiple cohorts and institutions
- All contributing schools were based in Scotland, and care should be taken when generalising to other contexts
- The study methodology cannot fully explain the mechanisms behind such attainment gaps

Introduction

Promoting fairness in assessment is a key priority. Success in medicine should be determined by ability rather than background characteristics like ethnicity, sex, or socioeconomic status. (1) There is an increasing emphasis on educational processes being “fair” to candidates of diverse backgrounds: besides the legal and regulatory requirements (2) there is growing acceptance that evaluating fairness should be a routine part of test construction and assessment. (3)

Despite this, candidates continue to experience different outcomes in medical education and training because they have characteristics which lead to them being treated differently by staff, students, and patients. The tendency for outcomes to vary in this fashion is usually termed Differential Attainment (DA). It influences every stage of medical education, and is a global phenomenon with similar problems manifesting in a range of contexts. (4,5) The varying treatment of some groups influences the likelihood of candidates completing medical school and affects selection methods. (6–8) Performance on measures of success at or just beyond graduation show a similar pattern (9,10) and, for example, ethnically white UK graduates are given higher marks than non-white UK graduates in postgraduate examinations with typically moderate (d = 0.22) effects. (11) After graduation, ethnically non-white and female doctors experience barriers to success on a range of professional and educational outcomes. (12–14) Students from underrepresented backgrounds are substantially less likely to be awarded high ratings from their clerkship directors, less likely to be given honours, and less likely to be given honour society membership. (15)

Such compelling evidence has led to calls to establish the mechanisms of DA, but this is challenging. Many historical assumptions – such as the idea that examiners are biased against some candidate groups – remain commonly cited despite evidence to the contrary. (16,17) Examiner bias does not appear to explain DA in postgraduate clinical examinations (18) or written assessment. (19) Qualitative research has emphasised a range of possible factors that can contribute to DA, including trust between trainers and trainees and the process by which those in difficulty are identified and referred to support networks. (20–22) Other research has suggested that unconscious biases may alter training pathways or assessment in the workplace. (4,13,23,24) Some authors now recommend a programmatic approach whereby each component of training is separately reviewed. (25)

As a result, evidence for the existence of DA is very strong but we have so far only a limited understanding of the mechanisms by which it operates or even whether DA increases or decreases with time spent in medical education. Compounding this, while a great deal of research has been carried out on access to medical school and postgraduate assessment, relatively little work has evaluated DA on assessment *during* medical school. In a large meta-analysis eleven of fourteen published studies examining undergraduate medical education used a single site and two of the remaining studies used only two sites. (11) Combined with the tendency to monitor attainment at only a single time point (typically finals) we know little of whether DA is of similar magnitude for different medical schools or remains stable during medical school.

This is an obvious limitation given the role of medical schools in providing the foundation of medical education and training. Due to the diversity of intakes, assessment choices, curriculum design and performance on postgraduate assessment (26,27) investigating DA at medical schools may help in several ways. By comparing different institutions, the effect of different recruitment strategies, curriculum types and policies on fairness in medical education can be explored. If the magnitude of DA is highly variable across institutions, it argues for a relatively larger role in medical school policy in creating DA. If DA remains consistent despite varying institutional contexts, it argues DA is either explained by factors outside of medical school control or that no current approaches are identifiably superior or inferior. By examining the data longitudinally, it becomes possible to explore whether DA increases or decreases over time. If DA is present from the earliest part of medical education, this suggests different mechanisms than if DA is minimally present at the beginning but then grows with time. Such work can therefore significantly improve medical education and support a fairer experience for doctors.

In this study we used data from four Scottish medical schools operating within a common regulatory framework. Our aim was to evaluate longitudinal DA across undergraduate medical education in 1,512 medical students, exploring disability status, domicile, ethnicity, and gender. Here we report on the longitudinal effects of DA for these groups and the impact of DA on student rank.

Methods

Participants

Participants were undergraduate medical students who had attempted (but not necessarily passed) a major written (multiple choice question) assessment near the end of medical school. All institutions operated under the UK medical education system (2) and new graduates typically embarked on a two-year foundation training programme as a doctor.

In total, 1,512 medical students were eligible for inclusion in the study. To be eligible, a student had to (a) have attempted (but not necessarily passed) the final written assessment, (b) have made the attempt by the end of data collection and (c) have provided demographic information.

The 1,512 students represented 74% of all available participants within the period of this study. Excluded subjects were typically those who had exited medical school before final assessment, experienced an interruption of study, or those who had intercalated close to the end of the study period and so had not yet sat finals. Due to the complexity of discontinuation it is theoretically possible for a student to graduate up to nine years after starting a five-year programme, which makes confirmation of discontinuation challenging. Candidates who did not attempt final assessment prior to the end of the period of data collection are not included in any analyses presented here.

Table 1 summarises the partner schools, total sample sizes and assessments used. All schools offered five-year MBChBs. The first two years of each programme involved an introduction to the fundamentals of medicine, anatomy, social issues around healthcare, and working with peers. Each programme offered an opportunity to intercalate, whereby candidates spent an additional year studying a topic in greater depth before returning to the core programme. In the later years, candidates rotated through a series of clinical placements to develop the skills and knowledge necessary to work as a junior doctor.

In each school, candidates sat a written assessment at the end of their first year. These featured Multiple Choice Questions (MCQs) and for two schools, Short Answer Questions (SAQs). For each question candidates were presented with a scenario and question. For MCQs candidates selected the correct answer from a list, whereas for SAQs candidates provided a short, written answer. The assessment was blueprinted based on programme learning outcomes and standard set by experts familiar with the curriculum.

Near the end of medical school, candidates sat another written assessment. Three schools delivered this in final year, while one (the University of Aberdeen) delivered it at the very end of the prefinal year. The blueprinting and standard setting process was the same as in the early assessment.

In each case, the assessments acted as a progression barrier: candidates needed to achieve a satisfactory mark to progress to either second year or graduation. A review by the authors identified that although there were some variations in curricula and teaching methods there were no significant differences in content and structure of assessments between programmes that would impact cross-school comparisons of DA.

Table 2 describes the participants according to important demographic characteristics. We report whether the candidate did or did not have a known disability, where they were domiciled before starting medical school, their ethnicity, and their gender. All recorded data was self-reported. For ethnicity and domicile, we aggregate data across many sub-categories into broad groups such as “Scottish domicile” or “white.” While a more detailed breakdown would be helpful, the small numbers in many groups prohibit this. The demographic characteristics selected for study are based partly on the concept of a "protected characteristic" for which there is a legal obligation to promote equality within the UK (28), partly on demographic characteristics known to be important from past research, and partly on availability of data. To give two examples of data availability, marital status and sexual orientation had levels of missingness that were too high to achieve necessary levels of power. The four categories described here (known/no known disability, international, non-EU/Scottish domicile, non-white/white and female/male) represent all those selected for full analysis and all analyses have sufficient power to detect medium effects. We selected Scottish (as opposed to whole UK) domicile due to Scottish domiciled candidates having already experienced the Scottish legislative and educational framework and having selected a medical school relatively close to home. Furthermore, differences in the funding approach in Scotland compared to the rest of the UK made merging the two groups less defensible. Non-Scottish domiciled UK students were included in the other comparisons and so e.g. an English domiciled student who provided valid information on gender would have been reported for that analysis.

Socioeconomic status (SES) was recorded in the dataset in two forms. Firstly, candidates had the opportunity to list parental occupation. Over 90% of candidates did not fill this in. A second proxy for SES was candidate postcode, which can be converted into an index of multiple deprivation. (29) However, it was not possible to effectively compare Scottish, non-Scottish UK, and international measures of SES within a single dataset. As such we did not explore this covariate further in the present study.

[Insert Table 1 about here]

[Insert Table 2 about here]

Data Protection and ethics

This project represented a considerable challenge under data protection legislation and required a careful and thorough evaluation of ethical issues. To ensure data protection, a designated team member undertook an honorary contract with each partner and worked in tandem with a data custodian at that school. This meant individualised data was never transferred outside of the school servers, and a thorough anonymisation protocol was used to verify that no "unique" combinations could identify candidates from their data patterns. Ethical approval was granted by the ethics committee for the College of Medicine and Veterinary Medicine at the University of Edinburgh (reference: 2018/7), and then separately approved by an ethics board and a data protection officer at each of the other schools. All participants gave informed consent. Prior to data analysis, all partners agreed to disseminate the results in public and to representatives of the study population: in this case, medical student organisations.

When describing inequities researchers must ensure individuals are described fairly and appropriately, without discriminatory language. Throughout this paper, we have used language which shows that group membership itself does not *cause* an attainment gap and is never a direct determinant of performance, and instead likely reflects systemic societal issues. We have provided some additional references which may be helpful in exploring language choice when describing historically under-represented groups. (4,20)

Patient and public involvement

The study was carried out exclusively on medical students and did not involve patients in any way. As such, there was no patient or public involvement.

Statistical analyses

Each medical school has a locally designed curriculum and assessment environment. We investigate written assessment as the most comparable form of assessment, as the available clinical examinations vary considerably across the schools in both timing and format. To allow like-for-like comparisons across different written assessments we converted each cohort of data to z-scores. (30)

A z-score is a standardised measurement, where a score of zero indicates the candidate has received exactly the mean mark on the assessment, and a score of +/ - 1 indicates they have received a mark one standard deviation above or below the mean respectively. This is analytically helpful because it allows for comparisons where relative (rather than absolute) differences are important. If a candidate from one medical school receives a mark of 75, and a candidate from another medical school receives a mark of 70 on two different assessments, it is difficult to know who is more capable. But if the z-score for each candidate is 0, this indicates they are of the same level of ability *relative to their peers* and that they are both average.

We used the Shapiro-Wilk test to model residual values to test for normality. (31) Although the normality parameters were violated ($W = 0.99$, $p < 0.001$) further investigation suggested that parametric testing would still be more appropriate as parametric tests are more effective at minimising the risk of false positives where the group sample sizes and standard deviations vary across groups. (32) Sample sizes were sufficient to detect small effects at 80% power for ethnicity, gender and domicile, whereas for disability status the unequal group sizes and small numbers of students self-reporting a disability allowed for only medium effects at 80% power. (33) Due to the low sample sizes *within* each medical school it was not feasible to compare inter-medical school variability with sufficient power. Likewise, it was not possible to compare intersectional DA (e.g., ethnicity *and* gender). We used Welch's t-test for significance testing as a more robust alternative to other t-tests. (34) All analyses were carried out using *R*. (35)

Design choices

We made several design choices that influence the final dataset. Most importantly, by only including candidates who reach final assessment we exclude the majority of those who experienced major difficulties early in their studies. However, the only alternative is to either measure graduation rates, which prevents granular analyses as the overwhelming majority of students pass medical school (36), or attempt some form of imputation to estimate final performance of candidates who never reached that stage of education, with significant uncertainty over the accuracy of such estimates. We opt for a simple approach of reporting data only where fully available. One consequence of this is that variability is higher in final assessment than in first year, with more candidates performing poorly, so most z-score change values were negative. For example, it would be possible for a candidate to receive an A in first year and an F in final year and participate in our study, but it would not be possible for the reverse to be true – unless the student successfully resat assessment *and* then completed within the specified timeframe. This can be considered a form of "survival bias" and approaches to the problem always require trade-offs. (37)

To investigate survival bias, we compared the ratios of those who did versus those who did not provide final year assessment results for each group. For example, we compared the ratio of non-white/white completers against non-white/white non-completers. No differences in the ratios were detected for any studied group. This likely reflects the fact that non-completion (by the end of the present study) was due to a variety of factors and did not in itself indicate academic difficulty.

Following this, we carried out a number of comparisons. Firstly, we calculated the z-score for each student in their first year and then final assessment. We explored the equivalence of school. We compared z-score *change* between groups to see whether attainment gaps were growing or shrinking during medical school. Finally, we ranked all candidates to see who would appear in either the top or bottom decile for the final assessment.

Results

We first tested whether the performance profiles of each school were sufficiently similar to pool data into a single sample. We compared the shapes of the distributions, frequencies of outliers, and the overall variability of each cohort. After confirming the equivalence of the cohorts, we pooled all data into a combined sample of 1,512 students.

Table 3 provides a summary of (a) the z-score for each demographic characteristic per assessment, (b) the relative change in z-score over time and (c) whether the z-score change *between* groups is significant. For the present study, we are not interested in the attainment gap at either the start or end of medical school – but whether the magnitude of the gap changes over time. We found that the gap grew significantly for white vs. non-white students ($t(449.39) = 7.37$, $p = .001$, $d = .49$, 95% Confidence Interval 0.34 to 0.58), for internationally domiciled vs. Scottish domiciled students ($t(205.8) = -7$, $p = 0.01$, $d = 0.61$, -0.75 to -0.42), and for male vs. female students ($t(1336.68) = 3.54$, $p = 0.01$, $d = 0.19$, 0.08 to 0.27). No significant differences were observed for candidates with vs. without a known disability.

[Insert Table 3 about here]

For the three significant analyses, non-white, internationally domiciled, and male candidates were awarded a relatively higher score at the start of medical school. By the end of medical school, they were respectively awarded a lower score then white, Scottish-domiciled, and female students. The effect size was medium when testing ethnicity and domicile, and small for testing gender. In summary, non-white, internationally domiciled, and male students experienced a relative decline in their achieved marks at medical school which cannot be explained by low attainment before or in the first year of medical school.

Finally, we estimated how often medical students of different demographics would appear in the top and bottom decile based on their z-scores vs. their expected frequencies based purely on how many existed in each category. Table 4 summarises the details.

[Insert Table 4 about here]

Decile 1 is the highest scoring decile and decile 10 is the lowest scoring decile. Students with a known disability, Scottish students and non-white students are over-represented in the bottom decile and under-represented in the top decile. Students with no known disability, and white students, are over-represented in the top decile and under-represented in the bottom decile. International students, and male students, are over-represented in *both* the top and bottom decile. Female students are under-represented in the top and bottom decile.

This analysis shows that many groups exhibit differences not just in mean performance, but also in variability, with some candidates being under- and over- represented at the extremes of the distribution.

Discussion

Statement of principal findings

Differential Attainment exists within Scottish medical schools, with small to medium effects. The analysis described here demonstrates both the considerable difficulty in organising datasets to longitudinally investigate DA, and the ongoing importance of such work. Even among successful medical students – and the overwhelming majority of those described in the present dataset have become doctors – DA exists. The fact that many attainment gaps grow during medical school suggests educational factors within medical schools may promote DA.

Strengths and weaknesses of the study

It is important not to over-state the findings. Small to medium effect sizes are consequential and impact student education, but there remains considerable variance between students of all groups. In this dataset, candidates across the attainment continuum were present in every group. In addition, the core purpose of medical education – graduating a safe doctor – has been met for almost all participants in the dataset. The gaps observed here must be placed in this context. Finally, as until now we have operated in an environment with almost no published data, there is a risk that organisations which attempt to directly engage with the problem of DA are criticised for the differences they reveal, which may in turn drive reluctance to explore the issue in depth. It is important that stakeholders support the exploration of DA across the sector.

This study represents a novel attempt to understand DA not as a fixed factor, but as a changing influence on student performance and behaviour. The sample size and range suggest we can be confident the findings are potentially generalisable to other UK medical schools. By opting for a straightforward methodology, we believe the findings are robust and can inform future policy.

Despite this, there are limitations. The challenges of organising a longitudinal study using data from a range of institutions with varying outcome measures should not be understated. We have made design choices – such as excluding those who failed before reaching finals – which may influence the pattern of results. Due to the relatively small sample sizes of some groups, it was not possible to explore “intersectional” DA for e.g. candidates who were non-white and female. (38) Due to the nature of the available data on SES, we were not able to include SES as a covariate in the present study. All candidate demographics were self-reported, and so some information could theoretically be inaccurate. While we consider the curricula and assessment of the institutions to be sufficiently similar to allow for a combined analysis, it is possible that local factors may have created some unidentified sources of variance.

The lack of a shared, standardised assessment across schools required the use of z-scores (or an equivalent method) and the presence of a standardised assessment, such as the forthcoming UK Medical Licensing Assessment, would have greatly simplified the analysis. (39)

Data collection was challenging, and it was clear that there was no expectation during data creation that assessment-level data would be required five or ten years after the assessment was sat. Medical education data should be thought of as "perishable" – it is possible that even relatively recent data is being lost, overwritten, or rendered inaccessible. If medical educators wish to investigate DA across time it is critical that better data collection practices are implemented, and historic data sources should be secured and documented in national-level databases. (40) The alternative is that we may establish excellent prospective analyses for which we will have no useful data for up to a decade.

Comparison with other studies and unanswered questions

DA exists across medical education systems across the world and should always be considered when designing teaching and assessment. (4,5) Our findings support and extend past work exploring DA in postgraduate medical education. (9,12,13,21) and at medical school. (15,24) Importantly, our study also confirms that we remain unclear, as a sector, on the mechanisms behind DA. (18,19) All organisations involved in medical education must proactively consider how they approach fairness in medical education, and evaluate the impact of DA.

The limitations described above are logical opportunities for future work. Exploring the impact of SES, analysing intersectional characteristics, and studying those who do not graduate may offer insights into both the scope and mechanisms of DA. Exploring candidate domicile in a more granular fashion (such as measuring the distance between home and their selected medical school) may be helpful, especially alongside measurements of SES. Importantly, the design challenges highlighted here will persist until institutions develop rigorous frameworks to investigate long-term changes in student performance.

Implications and conclusions

The present study demonstrates DA changes in magnitude during undergraduate medical education. Combined with evidence that candidates of some groups are less likely to be given awards (15) and more likely to experience prejudice (24), it is very plausible that some of the mechanisms of DA are located in, or caused by, aspects of medical education within medical schools. As such, institutions must consider the possibility that their actions contribute to DA and develop appropriate policies for investigation and correction. (14)

References

1. Tsouroufli M, Malcolm I. Equality, diversity and fairness in medical education: international perspectives. Med Educ. 2015 Jan 1;49(1):4–6.

2. General Medical Council. Outcomes for Graduates. Manchester: General Medical Council; 2015.

3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: AERA; 2014.

4. Murphy M. Teaching and Learning About Sexual Diversity Within Medical Education: the Promises and Pitfalls of the Informal Curriculum. Sex Res Soc Policy. 2019 Mar;16(1):84–99.

5. Prideaux D, Roberts C, Eva K, Centeno A, Mccrorie P, Mcmanus C, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011;33(3):215–23.

6. O'Neill L, Hartvigsen J, Wallstedt B, Korsholm L, Eika B. Medical school dropout - testing at admission versus selection by highest grades as predictors. Medical Education. 2011;45:1111–20.

7. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Medical Education. 2016;50(1):36–60.

8. Cliffordson C. Selection Effects on Applications and Admissions to Medical Education with Regular and Step-Wise Admission Procedures. Scandinavian Journal of Educational Research. 2006 Sep 1;50(4):463–82.

9. MacKenzie RK, Cleland JA, Ayansina D, Nicholson S. Does the UKCAT predict performance on exit from medical school? A national cohort study. BMJ Open. 2016 Oct 1;6(10):e011313.

10. Pershing S, Co JPT, Katznelson L. The New USMLE Step 1 Paradigm: An Opportunity to Cultivate Diversity of Excellence. Academic Medicine. 2020 Sep 1;95(9):1325–8.

11. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ. 2011 Mar 8;342:d901.

12. Linton S. Taking the difference out of attainment. BMJ. 2020;368:m438.

13. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, et al. Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature. Journal of General Internal Medicine. 2019 May 1;34(5):712–9.

14. Ufomata E, Merriam S, Puri A, Lupton K, LeFrancois D, Jones D, et al. A Policy Statement of the Society of General Internal Medicine on Tackling Racism in Medical Education: Reflections on the Past and a Call to Action for the Future. J GEN INTERN MED [Internet]. 2021 Jan 22 [cited 2021 Mar 8]; Available from: https://doi.org/10.1007/s11606-020-06445-2

15. Teherani A, Hauer KE, Fernandez A, King TEJ, Lucey C. How Small Differences in Assessed Clinical Performance Amplify to Large Differences in Grades and Awards: A Cascade With

Serious Consequences for Students Underrepresented in Medicine. Academic Medicine. 2018 Sep;93(9):1286–92.

16. Woolf K. Differential attainment in medical education and training. BMJ. 2020;368:m339.

17. Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. BMC medicine. 2017;15(1):1–11.

18. The Queen on the application of Bapio Action Ltd [Cliamant] v Royal College of General Practitioners [First Defendant] and General Medical Council [Second Defendant], in the High Court of Justice, Queen's Bench Division, The Administrative Court. 10th April 2014. EWHC 1416 (Admin) 2014, Available at http://www.rcgp.org.uk/news/2014/may/~/media/Files/News/Judicial-Review-Judgment-14-April-2014.ashx.

19. Hope D, Adamson K, McManus IC, Chis L, Elder AT. Using Differential Item Functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. BMC Medical Education. 2018;18(1):64.

20. Cleland J, Palma TF. "Aspirations of people who come from state education are different": how language reflects social exclusion in medical education. Advances in Health Sciences Education. 2018;23(3):513–31.

21. Woolf K, Rich A, Viney R, Needleman S, Griffin A. Perceived causes of differential attainment in UK postgraduate medical training: a national qualitative study. BMJ Open. 2016;6(11):e013429.

22. Atewologun D, Kline R, Ochieng M. Fair to refer? Reducing disproportionality in fitness to practise concerns reported to the GMC [Internet]. 2019. Available from: https://www.gmc-uk.org/about/what-we-do-and-why/data-and-research/research-and-insight-archive/fair-to-refer

23. Kristoffersson E, Diderichsen S, Verdonk P, Lagro-Janssen T, Hamberg K, Andersson J. To select or be selected – gendered experiences in clinical training affect medical students' specialty preferences. BMC Medical Education. 2018;18(1):268.

24. Cheng L-F, Yang H-C. Learning about gender on campus: an analysis of the hidden curriculum for medical students. Medical Education. 2015;49(3):321–31.

25. Karani R, Varpio L, May W, Horsley T, Chenault J, Miller KH, et al. Commentary: Racism and Bias in Health Professions Education: How Educators, Faculty Developers, and Researchers Can Make a Difference. Academic Medicine. 2017 Nov;92(11S):S1.

26. McManus IC, Elder A, de Champlain A, Dacre J, Mollon J, Chis L. Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. BMC Medicine. 2008;6:5.

27. Devine OP, Harborne AC, McManus IC. Assessment at UK medical schools varies substantially in volume, type and intensity and correlates with postgraduate attainment. BMC medical education. 2015;15(1):146.

28. Davies C, Ferreira N, Morris A, Morris D. The Equality Act 2010: Five years on. International Journal of Discrimination and the Law. 2016;16(2–3):61–5.

29. Noble M, Wright G, Smith G, Dibben C. Measuring Multiple Deprivation at the Small-Area Level. Environ Plan A. 2006 Jan 1;38(1):169–85.

30. Abdi H. Z-scores. In: Encyclopedia of measurement and statistics. Thousand Oaks (CA): Sage; 2007. p. 1055–8.

31. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3–4):591–611.

32. Zimmerman DW. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. The Journal of Experimental Education. 1998;67(1):55–68.

33. Cohen J. A Power Primer. Quantitative Methods in Psychology. 1992;112:155–9.

34. Delacre M, Lakens D, Leys C. Why psychologists should by default use Welch's t-test instead of Student's t-test. International Review of Social Psychology. 2017;30(1).

35. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996;5:299–314.

36. Arulampalam W, Naylor RA, Smith JP. A hazard model of the probability of medical school drop-out in the UK. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2004;167:157–78.

37. Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival Bias Associated with Time-to-Treatment Initiation in Drug Effectiveness Evaluation: A Comparison of Methods. Am J Epidemiol. 2005;162(10):1016–23.

38. Morrison N, Chimkupete P. Double jeopardy: black and female in medicine. The Clinical Teacher. 2020;17(5):566–8.

39. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Bere SR de. The medical licensing examination debate. Regulation & Governance. 2017;11(3):315–22.

40. Dowell J, Cleland J, Fitzpatrick S, McManus C, Nicholson S, Oppé T, et al. The UK medical education database (UKMED) what is it? Why and how might you use it? BMC medical education. 2018;18(1):1–8.

Table 1: Participants, data ranges and assessments used

School name	Sample size	Data range	First year assessment	Final assessment
University of Aberdeen	104	2014/2017	MCQ and SAQ	MCQ and SAQ
University of Dundee	202	2013/2016 & 2014/2017	MCQ	MCQ
University of Edinburgh	871	2009/2013, 2010/2014, 2011/2015, 2012/2016 & 2013/2017	MCQ and SAQ	MCQ
University of Glasgow	335	2014/2018 & 2015/2019	MCQ	MCQ

Note: Data range described the first/final year of assessment data for each cohort. "Multiple Choice Questions" (MCQs) requires students to select the correct answer from a series of options. "Short Answer Questions" (SAQs) require students to type or write a short answer. All assessments were written rather than clinical.

Table 2: Demographic characteristics of the study sample

Demographic Characteristic	Category	Institution	n	Total n	
Disability	Known disability	Aberdeen	13	102	1512
		Dundee	13		
		Edinburgh	74		
		Glasgow	2		
	No known disability	Aberdeen	91	1410	
		Dundee	189		
		Edinburgh	797		
		Glasgow	333		
Domicile	EU (non-UK)	Aberdeen	2	44	1512
		Dundee	17		
		Edinburgh	14		
		Glasgow	11		
	International	Aberdeen	9	146	
		Dundee	12		
		Edinburgh	88		
		Glasgow	37		
	Rest of UK	Aberdeen	24	500	
		Dundee	40		
		Edinburgh	354		
		Glasgow	82		
	Scotland	Aberdeen	69	822	
		Dundee	133		
		Edinburgh	415		
		Glasgow	205		
Ethnicity	Non-white	Aberdeen	27	298	1512
		Dundee	21		
		Edinburgh	157		
		Glasgow	93		
	White	Aberdeen	77	1143	
		Dundee	165		
		Edinburgh	665		
		Glasgow	236		
	Unknown	Dundee	16	71	
		Edinburgh	49		
		Glasgow	6		
Gender	Female	Aberdeen	67	877	1512
		Dundee	129		
		Edinburgh	480		
		Glasgow	201		
	Male	Aberdeen	37	635	

Dundee	73
Edinburgh	391
Glasgow	134

Note: Candidates of "unknown" ethnicity, "EU (non-UK)" and "Rest of UK" domicile students are not included in any analyses described in the present study. All demographic characteristics relied on self-report data.

Table 3: Z-score change during medical school study

Demographic Characteristic	Category	First assessment (mean)	First assessment (SD)	Final assessment (mean)	Final assessment (SD)	Change (mean)	Change (SD)	Significance/CI
Disability	Known disability	-0.15	0.94	-0.38	0.73	-0.18	0.93	
	No known disability	0.09	0.89	-0.05	0.93	-0.1	0.95	
Domicile	International	0.46	0.83	-0.4	0.92	-0.57	0.92	* (-0.75 to -0.42)
	Scotland	-0.08	0.91	-0.05	0.9	0.01	0.97	
Ethnicity	Non-white	0.15	0.93	-0.34	1.06	-0.45	0.96	* (0.34 to 0.58)
	White	0.04	0.89	-0.02	0.88	0	0.92	
Gender	Female	0.03	0.89	-0.01	0.88	-0.03	0.93	* (0.08 to 0.27)
	Male	0.14	0.89	-0.2	0.98	-0.2	0.97	

Note: * indicates statistical significance at p = .001. 95% confidence intervals are given for significant results. For model values see text. Statistical significance indicates the relative attainment gap between categories changed significantly during the course of study.

Table 4: Rankings of top and bottom decile by demographic characteristic

Demographic characteristic	n	Category	n category	Percentage	Decile	n in decile	expected percentage	actual percentage
Disability	1512	Known disability	102	6.75	1	5	0.68	0.33
					10	14		0.93
		No known disability	1410	93.25	1	145	9.32	9.59
					10	136		8.99
Domicile	968	International	146	9.66	1	21	0.97	1.39
					10	19		1.26
		Scotland	822	54.37	1	78	5.44	5.16
					10	85		5.62
Ethnicity	1441	Non-white	298	19.71	1	24	1.97	1.59
					10	54		3.57
		White	1143	75.6	1	115	7.56	7.61
					10	92		6.08
Gender	1512	Female	877	58	1	85	5.8	5.62
					10	81		5.36
		Male	635	42	1	65	4.2	4.3
					10	69		4.56

Note: n indicates the total sample size for that characteristic, while n category indicates the sample size for the individual category. Percentage indicates the proportion of students from that category in the overall sample. Decile 1 is the highest (i.e. best performing) decile, decile 10 is the lowest (i.e. worst performing) decile. N in decile gives the number of candidates who actually appeared in that decile, and the difference between the expected and actual percentage shows whether the category is over- or under- represented.

institutions must consider the possibility that their actions contribute to DA and develop appropriate policies for investigation and correction**STROBE 2007 (v4) checklist of items to be included in reports of observational studies in epidemiology***

Checklist for cohort, case-control, and cross-sectional studies (combined)

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1/1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	5/70
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	7/109
Objectives	3	State specific objectives, including any pre-specified hypotheses	8/158
Methods			
Study design	4	Present key elements of study design early in the paper	8/164
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	8/164 – 11/292
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	8/164
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	9/200
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	11/246 – 12 /268
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	9/179
Bias	9	Describe any efforts to address potential sources of bias	9/195 & 10/226
Study size	10	Explain how the study size was arrived at	8/168
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9/203 & 11/246
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	11/246
		(*b*) Describe any methods used to examine subgroups and interactions	12/270
		(*c*) Explain how missing data were addressed	12/283
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed	12/283

		Case-control study—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	
		(*e*) Describe any sensitivity analyses	NA
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8/164 & 12/294
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	9/200
		(b) Indicate number of participants with missing data for each variable of interest	9/200
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	9/179
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	NA
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	9/200
		Cross-sectional study—Report numbers of outcome events or summary measures	NA
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	NA
		(*b*) Report category boundaries when continuous variables were categorized	NA
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	12/294
Discussion			
Key results	18	Summarise key results with reference to study objectives	14/330
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	14/350
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	15/385
Generalisability	21	Discuss the generalisability (external validity) of the study results	15/371
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	3/57

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.