Evaluating the impact of interventions aimed at addressing variation in progression associated with protected characteristics known as ‘Differential Attainment’

Final Report

Submitted 13th August 2018
1 Executive Summary

1. The overall aim of this research is to enhance fair training for all. It aims to support and encourage training organisations and educators to evaluate interventions and programmes of support, and to share their learning with others in order to build an evidence base. It responds to the findings of previous research into Fair Training Pathways\(^1\) which identified that many organisations were delivering interventions aimed at improving fairness, but very few had evaluated the impact.

2. The objectives of this research were:
   - To identify and critically assess a range of measures that may i) indicate variations in medical education training pathways ii) help to quantify the impact of interventions aimed at reducing the variations associated with protected characteristics and improving fairness.
   - To produce practical guidance to support organisations in evaluating interventions aimed at mitigating unfairness.

3. The research used a mixed-methods research design, combining a realist literature review with individual semi-structured interviews and focus groups. As part of the literature review, 137 articles and documents were reviewed and a total of 76 potential measures were identified. A total of 46 individuals participated in the interviews representing a variety of stakeholders and a total of 15 individuals took part in four focus groups, including trainees, experts in differential attainment (DA) and psychometricians.

4. Following a process of critical appraisal and review, a total of 69 measures were included in the final framework. The identification of 69 potential measures is very encouraging, suggesting there are a breadth of options available to investigate the impact of an intervention or programme of work, providing flexibility and depth to any evaluation design.

5. The measures are categorised based on a post-hoc categorisation process (i.e. broad themes identified after collation). Nine categories were defined and these are outlined in Figure 1 below. Six of the categories relate specifically to medical training whilst the remaining three categories relate to desired evaluation outcomes (learning and knowledge, behavioural and motivational and affect related measures based upon Kraiger, Ford and Salas’ (1993) evaluation model).

---

6. Below provides a summary of the categorisation of measures, including examples and key messages.

**Selection Data**

- This section contains **12 individual measures**, which are all quantitative, already established and nationally available, although some would reflect local variation (i.e. undergraduate interviews). These measures generally demonstrate good levels of reliability, validity and standardisation. Some evidence of differential attainment is observed in these measures.

- These measures cover a range of stages in the training pathway: school educational attainment (A-levels), selection into medical school (UKCAT), entry into Foundation (EPM, SJT), entry into Specialty (MRSA, interviews).

- Two measures (A-levels and UKCAT) were identified solely through the literature review and 10 were identified through interviews or other sources.

- **Key Message**: Predominately the measures identified in the selection category would be most appropriate for benchmarking rather than evaluation of an intervention, given their timing in the pathway (especially at an undergraduate level).

- **Key Message**: As all measures within this category are quantitative, there is a need to consider the minimum sample sizes required for each group being compared. Using samples of an appropriate size will enable robust analysis to take place and create confidence that any results are not due to chance. Statistical expertise should be sought to define the minimum sample size and supporting factors, if required.
Examinations & Assessment

- This section contains 5 individual measures, which are all quantitative. Three are nationally available measures (ARCP, Work Place Based Assessment and Specialty Examinations), and these measures generally demonstrate good levels of reliability, validity and standardisation. Two further measures are available locally and are more variable in format and content (formative and summative assessments). Sub-categories within the examination measure were also identified, e.g. pass rates, number of re-sits and total score.
- Four of these were identified initially through the literature review and Work Place Based Assessment was identified through the interviews.
- The majority of the specific examinations identified took place at the postgraduate level; the broader assessments (e.g. summative) span the whole career pathway.
- **Key Message:** Specialty examinations are deemed to be most useful as a long-term evaluation measure because, while they are a measure of variation in progression, there is likely to be a time-lag between an intervention and access to examination results. This means that any evaluation could identify more embedded change as opposed to immediate, short-lasting reactions.

Training Progression

- This section contains 9 individual measures, which are all quantitative. These measures cover a range of training progression measures including completion of training stage, additional training time and access to additional learning support.
- The majority (7) are established measures available nationally and the existing data is expected to be reliable. Two of the measures (access to additional learning support, and formal remediation/disciplinary process) would need to be locally collected.
- Six of these measures were initially identified through the literature review and the remaining three identified through interviews or other sources.
- **Key Message:** This category contains some measures that tend to be at the end of the pathway (e.g. entry on the GMC specialist or GP register) and thus could only be used for long-term evaluation some years on from the intervention. Therefore, these measures may be less relevant in this context.

Trainee Experience

- This section contains 9 individual measures and cover a range of measures including trainee/trainer relationships, access to educational opportunities and relationship with institution.
- The majority of the measures would need to be collected locally, although some data is available through the National Training Survey and the Workforce Race Equality Standards (WRES) data.
- Eight of these were initially identified through the literature review, and 1 (frequency of trainer-trainee interactions) was identified through the interviews.
Key Message: Some measures within this category could utilize a self-report methodology (individuals complete a scale or questionnaire rating themselves on a series of questions). Self-report measures have a number of advantages; in particular they may be viewed positively by the trainees as this type of assessment gives them a sense of ownership over their own assessment. For example, self-report measures may reduce feelings of stigmatisation, compared to sole use of an exam or selection score without additional contextual information. A limitation of self-report measures is the variable ability of an individual to accurately assess a change in their own knowledge and behaviour, or the potential for individuals to respond in a not entirely truthful manner due to fear or lack of trust.

Trainer Experience

- This section contains 6 individual measures, including general knowledge of DA in the Deanery/LETB/trainer community and understanding how learning and training differs by country.
- All measures would need to be collected locally and could be collected using either a quantitative or qualitative format.
- One of the measures were initially identified through the literature review (trainers’ observations/perceptions) and the remaining five were identified through interview, expert stakeholder focus group or other sources.
- Key Message: The operationalisation of some of these measures is likely to depend on region and specialty.

The Environment

- This section contains 9 individual measures and includes institutional climate, availability and/or quality of additional learning support, and leadership diversity.
- Only one was identified through the literature review, and the remaining eight were identified through interviews or other sources, including four from the expert stakeholder focus group.
- The majority of these measures would need to be collected locally, with the exception of some data possibly available within the WRES dataset (leadership diversity).
- Key Message: The operationalisation of this group of measures is likely to depend on region and specialty.

Learning and Knowledge related measures

- This section contains 7 individual measures, including knowledge level of the multiple factors believed to contribute to DA, and demystification of assessments. All measures need to be collected locally.
- All seven measures were initially identified through the literature review but expanded or developed within the interview process.
- Key Message: Whilst these measures will need to be collected locally, examples of sample questions and/or validated tools are available and available in the Measures section within this report. Pre-validated scales or questionnaires may make the evaluation design more robust and can also save time and effort.
Behavioural related measures

- This section contains 2 individual measures: proactivity and self-regulated learning. Both would need to be collected locally.
- One was identified through the literature review (self-regulated learning) and the other was identified through interviews.

Motivational and Affect related measures

- This section contains 10 individual measures including engagement with learning, resilience and burnout. The majority of these would need to be collected locally.
- All 10 measures were initially identified through the literature review.
- Key Message: Documenting perceived negative attributes could be viewed negatively by trainees, due to concerns about what may happen with the data in future. Careful consideration around engagement with these measures is likely to be required.

7. Notably, in addition to measures relating to the trainee, measures relating to trainers and the wider environment are also included. It is recognised that differential attainment is the result of a variety of influencing factors, including the trainee, the trainer, the institution and the wider environment. As such, the identified measures are drawn from across these areas.

8. These measures have been classified into three categories of likely usefulness and/or appropriateness, helping guide an evaluator in their decision making whilst also taking into consideration their specific context. Information relating to measure availability/accessibility, robustness and considerations for use are presented within this report.

9. The measures, and information obtained from the stakeholder interviews and focus groups, have been incorporated into an Impact and Evaluation Guide. The purpose of the guide is to support and facilitate the evaluation of interventions run in medical education with the long-term aim of expanding the evidence base in this area. An increased evidence base for medical education interventions designed to address the causes of DA will support learning from others and increase awareness of successful interventions or programmes of work.

10. The guide is divided into three sections:

- An introduction to the evaluation of interventions or programmes of work in the context of DA. Here, themes arising from the research in relation to evaluating interventions in the context of DA are included i.e. challenges and mitigating actions relating to small sample sizes, how to categorise different group and addressing factors relating the multi-factorial nature of DA.
- An overview of how to conduct an evaluation, with links to example material.
- Examples of possible measures that could be utilised as part of an evaluation, including considerations to explore when using these.

11. The Impact and Evaluation guide will be made available on the GMC website.
2 Overview

2.1 Background and Context

2.1.1 The GMC sets the standards for medical education and training for doctors, which have ‘fairness’ as a core principle. The GMC aims to evaluate the fairness of training pathways as part of its quality assurance framework. Training should be designed and delivered to ensure all trainees have the opportunity to achieve their potential during their training. Progression should be based on ability, and individuals who are disadvantaged due to factors relating to protected characteristics should be supported to help them overcome barriers.

2.1.2 As part of a large-scale programme running from 2010, the GMC has worked with undergraduate and postgraduate medical educators and trainees to understand the experiences of doctors as they progress through the training pathway. A critical aspect of this work has been the exploration of variation in training outcomes experienced by groups of doctors who share protected characteristics, also known as ‘differential attainment’ (DA). The term DA is sometimes used as shorthand for the wider body of literature and phenomena that started as observed differences in attainment between groups with protected characteristics. Within this report, DA is defined as ‘variation in experience, perceptions or attainment for different groups with protected characteristics, which may or may not, be the result of unfairness.’

2.1.3 Data published by the GMC annually\(^2\) has established that similar attainment differentials exist across all specialty training programmes and across all regions and countries of the UK. Variation in attainment can be observed across groups when split by a number of protected characteristics, including age and gender. However, statistical modelling indicates the strongest effect is shown between different ethnic groups, and groups split by place of primary medical qualification (PMQ). As such, the agreed focus of this research is on the observed gap that exists between UK white and UK BME trainees, and the gap between those trainees who are UK trained and those who are not (international medical graduates: IMGs\(^3\)).

2.1.4 In order to address this variation, many educational institutions have developed interventions to support those groups identified as at risk. A 2015 literature review\(^4\) provided a summary of the support available for trainees and identified that no single intervention had been demonstrated to close the gap in attainment. “Fair Training Pathways for all: Part 2 (2017)\(^5\) also provided a summary of interventions as described by stakeholders in medical education but noted that evaluation of these interventions was extremely limited.

2.1.5 As such, the GMC now want to provide practical guidance and considerations to aid organisations, local educators, and trainers to evaluate any intervention designed to address the

---


\(^3\) IMGs are classified as a national of a country outside the UK, European Economic Area (EEA) or Switzerland who has graduated from a medical school outside the UK


causes of observed differences. It is hoped that high quality evaluations will be shared and may contribute to a more evidence-based strategy to address the causes of DA across the training pathway.

2.2 Research Aims

2.2.1 The overall aim of this research is to enhance the fairness of training by supporting training organisations and educators to evaluate the impact of interventions aimed at reducing disadvantages and improving fairness, and to share their learning about any effective approaches. The two key outputs are: An evaluation of available measures which can be used to explore the differential performance between demographic groups, and a practical guide for planning and implementing evaluations of interventions.

2.2.2 Using a mixed methodology, this research supports the completion of the following objective:

To identify and critically assess a range of measures that may indicate variations in medical education training pathways and help to quantify the impact of interventions aimed at reducing the variations and/or improving fairness.

2.2.3 It is acknowledged that measures for exploring the complexity of DA is not without its challenges. Whilst the outputs of these measures may provide some insight into the fairness of a training pathway (i.e. that variation could relate to something other than natural variation in capability and performance between individuals), definitive conclusions may not be possible.

2.2.4 Variation in training outcomes is one manifestation of differences in the experiences and perceptions of groups with protected characteristics. Therefore, this research explores potential measures of variability in these areas. However, variation in itself is not necessarily a negative concept; only when it is linked to differences in demographic groups that cannot be explained should this be a cause for concern.

2.2.5 It is acknowledged that there is limited definitive evidence in relation to potential measures, and a difference of opinion will exist in relation to the usefulness and suitability of the measures. This research project aims to collate all relevant information and provide an overview of the key findings and considerations. All measures will then need to be considered and interpreted in light of ones’ own context and locality and thus used on a contextualised basis.

2.2.6 The application of a qualitative research methodology generated outputs addressing objective 2:

To produce practical guidance to support organisations in evaluating interventions aimed at mitigating unfairness. This guidance should also help the GMC develop evaluative capacity in local teams and provide a benchmark to assess the quality of local interventions.

2.3 Outcomes of the Research

2.3.1 The research has produced two key products:

This report provides a:

- Summary of approach and research methodology
- Summary of results triangulation, demonstrating how the measures have been identified and how the various research outputs have been incorporated into the deliverables
• Overview of the key themes derived from the interviews and trainee focus groups
• Overview of the final measures and considerations related to their use
• Summary and conclusions

2.3.2 An Impact and Evaluation Guide has been developed to support and encourage those delivering interventions to include an evaluation of its impact, and this will be made available on the GMC website.
3. Methodology

3.1 Overview

3.1.1 This research used a mixed-methods research design, combining a realist literature review with individual semi-structured interviews (various stakeholders), and focus groups (trainees, experts in differential attainment and psychometricians). The use of these combined methodologies aimed to increase the validity and reliability of the outputs by enabling triangulation and validation of evidence gathered across the different methods. The realist literature review, the qualitative interviews and the focus groups supported completion of both research objectives. A full description of the research methodology is provided in Appendix A.

3.2 Realist Literature Review

3.2.1 The purpose of the literature review was the identification of possible measures of variability within the training pathway that may or may not indicate unfair training pathways drawing on literature describing both specific interventions, and medical education more broadly. A review of the organisational psychology literature was also undertaken to review models, measures and methods of evaluation. The purpose of the review was not to undertake an evaluation of the evidence related to any identified measures, but to identify a broad range of potential measures for inclusion in the Measures Classification Framework. The full search strategy and scope can be found in Appendix B.

3.3 Interviews

3.3.1 The primary purpose of the interviews was to identify measures of variability (which may or may not indicate unfairness) and seek insight into their efficacy as a means of evaluating the impact of interventions which seek to improve fairness (objective 1).

3.3.2 A secondary purpose was to explore potential issues or challenges to evaluation and related solutions. These interviews thus helped to inform relevant considerations when implementing an evaluation within this context and fed into objective 2; the Impact and Evaluation Guide.

3.3.3 Telephone interviews were conducted and typically lasted for 45-60 minutes. A total of 41 one-to-one interviews and one virtual focus group consisting of five individuals (46 individuals in total) took place. Table 1 below provides an overview of the region, specialty and training pathway point of interviewees.
<table>
<thead>
<tr>
<th>Region</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>34</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td>2</td>
</tr>
<tr>
<td>Scotland</td>
<td>3</td>
</tr>
<tr>
<td>Wales</td>
<td>7</td>
</tr>
<tr>
<td>Specialty/Area</td>
<td></td>
</tr>
<tr>
<td>GP</td>
<td>18</td>
</tr>
<tr>
<td>Physician</td>
<td>8</td>
</tr>
<tr>
<td>Anaesthetics</td>
<td>3</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>2</td>
</tr>
<tr>
<td>Emergency Medicine</td>
<td>1</td>
</tr>
<tr>
<td>Surgery</td>
<td>2</td>
</tr>
<tr>
<td>Paediatrics</td>
<td>1</td>
</tr>
<tr>
<td>Haematology</td>
<td>1</td>
</tr>
<tr>
<td>Other (E&amp;D/policy advisor, training support unit)</td>
<td>10</td>
</tr>
<tr>
<td>Pathway Point</td>
<td></td>
</tr>
<tr>
<td>Postgraduate</td>
<td>40</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: Breakdown of interviewees

3.4 **Focus Groups**

3.4.1 Three formats of focus groups were held; with trainees, expert stakeholders, and psychometricians, with the purpose of validating and triangulating the findings from the interviews. The focus groups took place once the preliminary Measures Classification Framework had been as fully populated as possible using data from the interviews/literature review.

3.4.2 Two trainee focus groups were held; one in London and one in Newcastle. A total of 9 trainees took part in the focus groups. Four experts took part in the expert stakeholder focus group. Two psychometricians took part in the psychometrician focus group.

3.5 **Qualitative Analysis**

3.5.1 Thematic content analysis was employed for the analysis of interview and focus group data. Overarching themes or structure codes relating to areas of interest as determined by the interview schedule were identified (i.e. general overview/understanding of differential attainment and how it is examined, examples of interventions and considerations, measures of variability) with additional layers of coding within these themes.

3.6 **Triangulation of Data**

3.6.1 **Objective 1**: Potentially relevant measures were captured in a preliminary Measures Classification Framework (see Results section for full details) as one output of the literature review. This information was then triangulated with information emerging from the interviews and focus groups to update and progress the framework. Additional measures were either
added into the framework, or evidence about a measure (including considerations for its use), were incorporated into the framework.

3.6.2 **Objective 2**: Themes and patterns in relation to broader areas relating to DA, interventions and evaluation were identified and triangulated between interviewees, focus group attendees, and the literature review. These were used to inform the content of the Impact and Evaluation Guide.
4. Results

4.1 Literature Review Output Summary

4.1.1 The purpose of the literature review was to identify and categorise potential measures of variability (which may indicate unfairness) as opposed to undertaking a review of the literature content. As such, a detailed review of the outputs of the literature review is not presented but a summary of broad findings is provided below.

4.1.2 A total of 125 articles and papers were included in the detailed review based on the search strategy. A further 12 were identified through the interview process or other sources and were subsequently reviewed.

4.1.3 As had initially been hypothesized, there was a fairly limited range of published data and literature pertaining specifically to the identification and measurement of variability, in relation to DA.

4.1.4 Regarding the grey literature, most references to differential attainment focused on highlighting its existence conceptually or anecdotally or referred to it in brief. Therefore, of the literature reviewed, only a small proportion of grey literature identified was included in the final detailed review. The majority of the grey literature included came from GMC reports, and others (a minority) from relevant organisations such as BMA, the Higher Education Funding Council for England, the NHS Professional Support and Development website and the Cabinet office.

4.1.5 Within the academic literature, there was fairly limited rigorous research on differential attainment in medicine. However, there were relevant papers and articles on assessment, interventions and performance differences. Journals which contained the largest proportion of relevant literature were Advances in Health Sciences Education (AHSE), BMC Medical Education, Teaching and Learning in Medicine, and British Journal of General Practice (BJGP).

4.1.6 Whilst the focus of the search was on ethnicity and PMQ, articles and papers related to group differences across a range of sub-groups, including ethnicity, gender, PMQ (i.e. IMG vs UK graduate), and socio-economic status (SES). A small number also referred to rural vs urban. This research was included as measures identified could be relevant to a number of sub-group types; that is, a study investigating differential attainment in gender groups may include measures and methodologies also applicable to a study investigating differences across ethnic group.

4.1.7 In terms of the type of intervention each measure related to, most of the literature focused on trainees. There was a smaller focus on measures focused on trainers, and less still on measures related to leadership or the wider environment.

4.1.8 With regards to exploration of the broader evaluation literature, most research related to evaluation frameworks, such as the Kirkpatrick model (1967) (reaction, learning, behaviour

---

and results) or Kraiger, Ford and Salas (1993)\(^7\) (cognitive, skill-based and affective outcome measures).

### 4.2 Interview and Focus Group Output Summary

4.2.1 Whilst the primary aim of the interviews and focus groups was to define and validate potential measures, they also generated information around challenges to evaluation and potential solutions which would feed into objective 2; the Impact and Evaluation guide in terms of supporting evaluators to effectively evaluate their interventions. The themes arising from the interviews have been analysed and summarised below\(^8\). Highlighted in bold is the theme and highlighted in blue is the implication for the current research.

#### Theme 1: Differential Attainment as a concept

There was a rising awareness identified through the interviews that the issue of DA is not just a phenomenon unique to one particular group of trainees: it does not simply pertain to the performance of doctors who are of a different ethnic origin or trained overseas but is a broader concept that encompasses a wide range of doctors and issues relating to the wider training environment. Therefore, it is important that measures utilised should encompass factors relating to the environment as well as outcomes and a wide range of demographics. As one stakeholder stated:

“I think that the vast majority of trainers and trainees also, when they think about it they think, ‘Oh, this is a subset issue, this is an issue for IMGs, I don’t really need to hear about this’... this is the biggest challenge I have had in my role I think is to actually persuade people that this is about training overall, it’s about quality of care overall, this is not just relating to IMGs.” (S\(^9\))

The value of having multiple measures that are disseminated was highlighted as it was noted DA is often not perceived as an issue, or is something that educators or others are not aware of, until data is presented.

“So everybody’s initial reaction to the question of differential attainment is, “Not in my patch.” Until you actually show the data and having done so everybody gets it.” (S)

This may be because people are potentially embarrassed as they feel that this somehow shows the school or institution is discriminatory, or that this issue feels simply too big to deal with. Fear of being unfairly accused of racism was also cited as a reason for DA not being openly discussed or a general sense of awkwardness and difficulty in talking about it.

“It’s not something that comes up that I’m aware of, certainly in my trust and we’re not untypical of other local providers, but perhaps just feel uncomfortable about discussing it.” (S)

Responses from trainees also alluded to a reluctance to discuss the issue, including words such as ‘stigma’, ‘taboo’ and ‘failure’. This was also referenced by the stakeholder group;

“There is a subsection of IMGs which feels very discriminated against and they can be quite vocal.” (S)

---


\(^8\) Coding for comments is as follows (S = stakeholder) (T = trainee)

\(^9\) Stakeholder
Whilst it is important to acknowledge these perceptions and views, and to accept that using data to show the differences may reinforce the perception of stigma, this does need to be balanced against the need to encourage open discussion and exploration of the issues. If the value and educational reasons for targeted interventions are clearly communicated, then this will be one step towards attempting to remove perceptions of stigmatisation.

There is also likely to be a lack of awareness about the issue within the trainee population which reinforces the requirement for improved communication about the issues; one trainee responded;

“This is the first time I’ve ever heard about this differential attainment and us being open about it. I don’t think we, as doctors, talk about these things really.” (T10)

However, for some it was clear that there is a sense of engagement with the issue;

“Everybody is quite willing to engage around it given that that’s what the data show... Because there’s so many stakeholders, and there is a discussion about it.” (S)

and it was noted that only with acceptance there is a problem, and a real engagement with DA, that progress can be made.

Theme 2: The availability of information through which to examine DA

A number of themes arose in relation to challenges associated with the availability of information through which to examine and research DA. This section outlines some of these challenges, and where appropriate, comments on some possible ways these can be mitigated.

One theme was that some institutions and educators may not have a clear idea of how trainees from different demographic groups are progressing through training;

‘We recognise that most schools do not know how many IMGs they have in their training group, they don’t know how many have failed, how many have passed. Most Training Programme Directors will not know [what is] the pass rate of their trainees.’ (S)

This will be compounded by the fact that trainees come in through different routes (i.e. Medical Training Initiative (MTI), and Locum Appointed for Services (LAS)) and their data may not be as easily accessible as those who come through more traditional standard core training routes;

‘There are multiple routes of entry into training programmes, so not everybody comes through the standard core training route, but that is the largest group. [For] That group the data is probably most easily available, because all Heads of School will get that data and they should be able to pull out the data, looking at people by their primary medical qualification, by any other particular characteristic, they should be able to do that. However, there will be other trainees who will come through different routes, some will come through MTI, some might come in as for LAS posts, others might have come through some other MTI route, so those numbers are really small, but they will still be there, so yeah, I think it requires a little bit of extra effort on the part of the school to get that information.’ (S)

---

10 Trainee
It is clear therefore that more can continue to be done to raise the awareness of the importance of accessing and utilising this data and supporting educators and others to be able to do this.

One common theme was in relation to the challenges associated with the small sample sizes that are available for some protected characteristic groups, especially once data is further separated by region or specialty. Small samples may not generate meaningful or reliable analysis and this issue is compounded by broader challenges with statistics i.e. the interpretation of findings can be subjective and inferences have to be made;

“So you can see that there is a significant difference in the data but then when you actually dig into it, it turns out that you’ve got three students who hit a particular demographic who have done particularly well or particularly badly. And that is your group size.... the thing that is important when doing that is having someone who can interpret it in a way that makes some sense rather than taking everything at face value. If we only pay attention to P values I think there is a loss of the quality of the data we have.” (S)

Ways that this can be overcome to help support educators to be able to make meaningful conclusions from their data in these circumstances include;

- Collaborating with individuals who possess some experience or expertise in the area of statistics.
- Examining multiple sources of data to provide a more complete picture.
- Utilising measures that do not require such large sample sizes. These include more qualitative and locally collected measures that are specific to the intervention aim and tailored to the anticipated outcomes that will still show a meaningful difference, even with a small sample. Outputs and conclusions generated from a smaller sample are likely to be indicative rather than definitive, but this is still a valuable addition to a growing evidence base.
- Examine opportunities to aggregate information over time, across several cohorts, or across institutions if the intervention is conducted in multiple places.
- Grouping together smaller demographic groups into broader categories.
- Investigating what data may be accessed through other organisations, especially if the data set has limitations or gaps. Institutions such as the GMC publish data and may also provide bespoke anonymised or aggregated data, linked to outcomes which may be useful for you. Research databases such as UKMED\(^\text{11}\) can allow data to be linked to others through a secure portal, and the results shared.

These challenges are compounded by the fact that performance or other data is not always tagged with demographic data. This is, to some extent, likely to be due to the voluntary provision of this information by trainees, and a reluctance to share this information. Ways that lack of access to demographic data can be overcome include;

- Sharing with trainees the purpose of collecting the data, and the broader benefits that this may bring, to encourage participation and engagement.
- Using pooled, anonymised research databases such as UKMED

\(^\text{11}\) https://www.uk-med.org/
• Using data which is published by others

A further challenge described was the way that different groups are categorised for the purposes of evaluation, either through existing published data or local decisions about how to categorise the data. Whilst using broader groups such as BME can increase sample sizes and provide a means for understanding broad trends, reducing individual experiences down to a single shared demographic can be limiting and may mask differences that exist within these groups. A need was recognised to define characteristics or background in more detail where this was possible and if not, to recognise this as a limitation when making conclusions.

“If you take something like ethnicity, white versus BME would give you one group analysis. White versus Asian versus others might give you another. So, some of it is very subjective into how you actually arrange the data.” (S)

As well as the difficulties with categorisation of groups from a methodological viewpoint, the potential concerns related to ill-defined labelling from a more theoretical standpoint was raised;

“Issues are much more subtle and more complex than grouping cohorts together under headings that are poorly defined... I see it as a much more individual issue, it’s far more complex.” (S)

“The overlap of race and ethnicity for example, whether people are clear about that and if you are multi-cultural (for want of a better term) which ethnicity would you put down, would that be misleading in the sense that you function effectively in several different spheres.”

However, whilst it is recognised that broad categories may be challenging and limited, they can still be helpful for starting a more nuanced discussion about individual experiences.

A final area raised was the stated challenges associated with examining DA as a multi-factorial issue. It is recognised that DA is the result of multiple influencing and confounding factors, including the trainee, the trainer, the institution and the wider environment. As there is no single cause of DA, attempts to measure it can be complex and can lead to a lack of certainty around which measures might be relevant. As such, attempts to reduce it may consider a range of individual or environmental factors. It is also challenging to confidently attribute any change in DA to the intervention itself. Change may be influenced by factors over which institutions have limited control (or no control) and have no or very limited data.

“One of the discussions that we have about differential attainment is [that] the issues are complex and they’re multifaceted, and probably for different groups and possibly even within those groups for different individuals there are different things feeding in and different factors potentially feeding in. And I think one of the complicated issues here is probably there are multi factors in play and ...there may not be consistent contributions even within groups” (S)

As evaluators, there are a number of ways these challenges could be overcome, and these include;

• Be clear about the outcomes sought as part of the evaluation. Clarifying this from the start will support a focused intervention and an evaluation that specifically targets these outcomes.
• When interpreting the results, consider the other variables that could be influencing the findings. Whilst it is unlikely to be feasible to measure or quantify the impact of these other variables, these should be acknowledged when reporting.

Despite the challenges outlined, it was clear from respondents that this did not mean DA was not important to address and that there was indeed benefit in doing so to explore and build the evidence base. While DA research and evaluation will remain challenging, contributing factors should be mitigated through a marginal gains approach and the results of an evaluation will still be valuable even if other factors may have had an effect on the reported findings.

“So what we’re saying is the factors that lead to differential attainment...are complex and varied, but that does not reduce the need to seek to understand and address those sorts of issues.” (S)

Theme 3: Intervention considerations

Whilst outside the direct scope of this research, information was gleaned about interventions and considerations when designing or running them. Some respondents had views on the challenges associated with implementing interventions. One challenge mentioned was the cost, resource and time taken to create and run interventions, and that this activity is often not seen as a priority.

“This is just how it is when it’s under pressure, you know, all the other bits. So, there’s a squeeze on the available time for educational issues whereas, in the past, people have maybe had more flexibility with their time. A lot of these things are, you know, it’s almost like goodwill.” (S)

This lack of resource and time can be directly aligned to the challenges associated with running evaluations. However, it is clear that some educators are managing to overcome these challenges, as some evaluations are taking place which appear to be due to combination of having the available resource, and the commitment, engagement and determination of the educator involved. It would be beneficial for those who are managing to conduct these evaluations to share learning on these tactics and strategies.

One challenge that emerged around the success of interventions was the need for acceptance that there may be an issue in the first place or that the intervention is appropriate. For example, there may be a perception of a ‘negative label’ attached to attending interventions, and some groups are less likely to ask for help or accept it when offered.

“So there is that thing, ‘I don’t want to be labelled as a person in need’. ” (S)

“I think some of the stigma needs to be taken away before it can really work, because otherwise people won’t ask for it.” (T)

Whilst during the interviews there was a view that targeted interventions may have their limitations, it is acknowledged that there is a difference of opinion around this concept and there are valid, evidence-based educational reasons for using a targeted approach and this should not be discouraged. Neither the less, suggestions were made relating to who interventions are made available to and whether a particular group should be targeted.
“What distresses me is the areas who give extra support to the bottom 10% of recruitment and selection, 90% of whom are IMGs. So you’re already beginning to label them as needing remedial treatment.” (S)

“And actually they decided that looking at an intervention that looked to make the situation better for everybody...and I think, until you’ve got a system whereby we aren’t isolating groups, and then causing issues, because we’re getting a huge backlash from trainee groups at the minute.” (S)

These quotes highlight the need to ensure that communication around any type of intervention and the resulting evaluation is clear and appropriate.

Linked to this point was consideration of the purpose of these interventions, and if there should be interventions specifically put in place to support specific demographic groups, or if training and support should be made available to anyone with an educational need, or to everyone equally regardless of their individual training needs;

“We say we want to raise the general standard of teaching, and you know what, by doing things like that we benefit much wider groups.” (S)

Many respondents who described their interventions stated they were not framed as ‘targeting differential attainment’ but focused towards any individual that may need a little more time and effort to get them through an exam. It is important to remember that doctors from all demographic groups may struggle with exams or other parts of training, and doctors from all groups may excel. One respondent framed this using the example of how an individual may not be in a particular group, but performance may be related to their current circumstances:

“People can move up and down so you know you can be doing very well at one point in your life and then circumstances change and there may be health problems or personal issues that impact on performance at work for a number of reasons. And those, if the individual is supported through those times, can result in them, with appropriate support, they can get back to functioning at the higher level that they were functioning at before.” (S)

How the intervention’s design or implementation is influenced by local context, such as the demographic breakdown of the trainee population or general population, was also discussed. It was considered important to take account of the broader context and remember that what may work in one area may not work in another.

“London might well be a very different world and things that you can do in London may turn out to be very different from things that you can do in Aberdeenshire.” (S)

However, there was a view that it would be beneficial to have a model that could be shared, adapted and implemented on a wider scale which supports the need for evaluations to be shared and discussed more widely so that others can benefit:

“It’s developing a model which is transferable across different trusts, different regions.” (S)

Rather than viewing interventions separately, it was also noted that interventions should be designed to work in a complementary way, rather than working in isolation.
“So that’s the idea behind those two interventions and I think that ultimately they’ll probably work together and probably work additively or cumulatively and that’s the idea.” [referring to an educational intervention for trainees designed around areas where international medical trainees tend not to do as well and developing the capacity in the educational supervisors to have increased understanding and increased awareness of the issues]. (S)

There was a clear enthusiasm and keenness to share learning but aligned with other research, it was clear that this is not shared currently in the majority of cases.

“Others I’m sure do great work, better than me, and no one knows about it because we’ve not had a forum or the ability to share that.” (S)

“Having something like a guidance on best practice or these are ideas that have worked in this area, you might like to think about it. I think that would be helpful because a lot of people won’t know where to start.” (S)

This perhaps links to a limited awareness of current forums for sharing e.g. Academy of Medical Educators and BMA have run conferences specifically on the subject and the GMC now has resources available to support with the sharing of learning.

Although there was evidence of some educators currently trying to do this:

“Our newsletter [which] goes to all trainees and all trainers. We’re going to start showcasing successes and we’ll put out some of the successes we showcase. It could be showing UK BMEs and IMGs and will ensure there is a real sharing of success stories that captures the depth of these across the whole groups.” (S)

**Theme 4: Evaluation of Interventions**

Whilst many interventions related to supporting trainees in need were discussed, there was limited evidence of any evaluation of impact taking place. This was predominantly due to a lack of time and resource to be able to conduct evaluations.

“The biggest problem we have is lack of resources in terms of evaluating all this, it’s not that we don’t want to do it, we just don’t have the admin staff.” (S)

Where evaluation has been taking place, this has predominantly been immediate feedback on the course, and anecdotal feedback;

“We’ve had anecdotal feedback, we get messages saying ‘Thank you so much’, or ‘I finally passed that exam’ or whatever, but we haven’t got any proper evaluation beyond the immediate course evaluation.” (S)

Although there was some evidence of attempts at follow up surveys and more qualitative evaluation, the difficulties with having an appropriate outcome measure were also raised, highlighting the importance of being able to access multiple and diverse measures;

“We could collect data, especially for the CSA and limit it to trainees who are potentially having difficulties or have already failed so we could look at outcomes from that. But that doesn’t have any specific language/culture content so you know it’s quite hard to see... it’s pretty difficult to put your finger on what is working, if anything.” (S)

A practical challenge of getting individuals to respond to surveys or questionnaires that are devised for such evaluations, and subsequent low response rates, also arose as a theme;

“The ones who haven’t passed are very, very reluctant to respond to any assessment questionnaire and that is to do with feelings of failure and you know not really feeling part of the system.” (S)

and suggestions were made as to how to overcome this:

“...because if you wait and then send them a questionnaire which is what you would normally do in a study it won’t work, you won’t get the feedback. So either you’ve got to do it there and then when the face to face intervention is happening or you’ve got to go again and have some other kind of meeting, you could build in for example you know a review meeting [of sort of] a month after the first one, how are things going? And get people to do the questionnaire there and then.” (S)

“Another approach might be to ring them ... if you rang them up...you’re likely to get better data because not only you’re going to talk to them, but you’d also be able to maybe ask the questions in such a way that a questionnaire wouldn’t capture properly.” (S)

“I find that that’s always quite helpful to try and bolt on sort of an extra sort of half an hour or something or over a lunchtime to obtain feedback which could be like a focus group or kind of qualitative interview which you could do within [sort of] individual trusts.” (S)

One of the challenges associated with evaluating interventions was that often multiple interventions are taking place at one time, so defining specifically what is making the difference can be challenging.

“...much more difficult to unearth the interventions that might help ... which are the ones that worked and [which are the ones that] missed it.” (S)

Controlling for multiple interventions or other factors that may be influencing the outcomes is unlikely to be feasible in this context, and thus this is something that needs to be acknowledged and considered when drawing conclusions.

This point links to the multi-factorial nature of DA, and the related point that any intervention would have difficulty addressing factors outside of the training pathway, such as geographical, societal or social influences. In addition, the majority of education takes place outside of the intervention. Thus, demonstrating that an intervention has made a difference may be very difficult.

“When we looked at the figures early on we were very despondent because it didn’t seem to be making much difference. And then we reflected and said ‘well, in a sense you wouldn’t expect one five day course to make much difference because 99% of the education is taking place in the practice’. And so that’s where the bulk of the change is going to be and unless you can contribute to that process.” (S)
This highlights the importance of ensuring that the appropriate measure, or combination of measures is used, considering short term versus long term measures and potentially using a combination of qualitative and quantitative measures.

**Theme 5: Existing GMC Measures**

As part of the research, there was an interest in understanding interview respondents’ awareness of currently available data to measure differences in experience and perception during training, i.e. examination data and data that is held at a national level by the GMC. Respondents were asked about their experiences of using the data, whether they found the data useful, and what changes or adjustments in data provision or presentation could be made that might increase its usefulness.

The majority of respondents had little insight or understanding concerning the data available from the GMC. Some were not aware of it at all, and others knew only that it existed. This may be in part due to the population of individuals involved in the research, especially at the latter stages of the research who were less likely to be leaders in medical education or have a direct connection with the GMC. As one respondent indicated;

‘Many people don’t look at it really, and they should, they should, you know. And that’s where the value is… and that’s why I believe that it’s good they collect it.’ (S)

There were a number of respondents (mostly those participants with more insight or expertise in relation to DA) who were aware of the GMC data and provided their views. Some respondents felt the current data presentation is adequate for current needs;

‘It’s really cut specifically for college exam outputs and for ARCP outcomes, and recently produced information relating to overall satisfaction. So it’s still down at the moment to these three key outputs…and in fairness that’s enough for the conversations to take place.’ (S)

Other indicated the drawbacks of not being able to break down on a regional basis;

‘So recently we had a workshop for educators around facilitating exam progression and I was able to access the GMC data and show the data for different groups; international medical graduates, UK BME and then UK White but you cannot break that data down on a regional basis. I suspect that there are good reasons for that in that you might access some data and maybe misinterpret it.’ (S)

Others suggested presenting the data by number of attempts/when they started the exam process;

‘Sometimes there are other things that you would like to know, in terms of maybe how many attempts someone has had at an exam before they have actually passed it, or you know when they started an exam process because the data you get is a snapshot in time and you’re not exactly sure as there might be subtleties around that snapshot in time.’ (S)

or using the data to explore trends over time;

‘Probably the trends over time would be helpful and to show if we are making significant improvements. I think for organisations like us and probably a lot of people are looking at
national policy we’d like to see the sort of the overall picture as well. That helps to, sort of, compare and contrast what’s going on.’ (S)

Other suggestions including information about first language;

‘They are from all over the world, other than the European Union. Now that’s just a broader brush to say that they are differentials. So it is a question of what is their first language. Is English in fact their second language, or third or fourth even at times for some of these people and it will be third to some of them. Have they studied in English and have they done the elementary, of course and will they have a very book orientated medical degree.’ (S)

The largest proportion of comments were in relation to providing information relating to previous attainment, to help understand how best to support trainees and share good practice.

‘I think that has to be seen in relation to say previous exam results in A Level standard and then track students over time because my understanding is, that it’s nothing to do with ability, this is about how information is absorbed. If there is a problem here then we have to be using data to get an understanding of what is happening.’ (S)

‘So you almost need to have some baseline information on people of what the situation is when they come into a training programme and what the situation is when they leave the training programme and sometimes, you are looking at the data, while it’s very useful it doesn’t necessarily answer that question for you.’ (S)

‘We do an awful lot for our international graduates and we think we are really good from that point of view. They just take more support but actually if you look nationally that value add[ed] from our programme isn’t represented anywhere. They seem to be proactively not released and that data isn’t shared. It’s undermined to some degree how much we do and how much effort we put into something and how much money we put into it supporting these trainees. If it can’t be sung about at a national level and if we can’t compare ourselves with each other how do we know which area we are going to learn from?’ (S)
4.3 Development of Preliminary Framework

4.3.1 As an output of the literature review, all potentially relevant measures of fairness were captured in a Measures Classification Framework. The definition of a measure used was;

“A metric that may be quantitative or qualitative in nature that could be used to identify variation in experience, perceptions or attainment for different groups with protected characteristics”

4.3.2 The Measures Classification Framework was used as a working document to capture all potential measures. The guiding principle of data gathering was to include everything that could be relevant and then filter this through a process of review and validation to reach a final set of measures. This information was then triangulated with information emerging from the interviews and focus groups and used to update and develop the framework. As such, the final framework is a triangulation of the findings from across the literature review, the interview and the focus groups stages.

4.3.3 For some of the measures, evidence from the literature relating to the outlined criteria was limited. Where this was the case, data was sought from further targeted research into these measures (i.e. internet searches relating to the Annual Review of Competence Progression (ARCP)) and these areas were also focused on within the interview stage to help understand the efficacy of these measures.

4.3.4 Criteria used to critically assess the salience and efficacy of the measures included:

- **Details of the measure;** pathway point, whether national or local, whether quantitative, qualitative or mixed, and what type of intervention the measure may be useful for
- **Data access and collection;** does it already exist and, if so, how can it be accessed. Level of completeness/standardisation
- **Evaluation criteria;** reliability, validity, statistical power\(^\text{14}\), differences in attainment evidenced (i.e. have there been significant differences found within the measure for different protected characteristic groups), and estimated perceptions of fairness to be used as a measure i.e. (through trainee focus groups, would these measures be perceived as fair to be used for this purpose). Some of these criteria are only relevant to existing measures, rather than measures that would need to be locally collected, as no data is available to inform these criteria.

4.4 Summary of Preliminary Framework

4.4.1 The preliminary framework contained a total of 114 potential measures of variability. 76 were identified as part of the literature review and a further 38 were identified through interviews, focus groups and other sources.

4.4.2 In terms of the pathway point, many of the measures related to postgraduate training and are available at a national level. There were very limited measures that are directly related to medical

---

\(^{14}\) Statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error (concluding there is no effect when there is one) goes down.
school and Foundation training, although many of those identified could be applied universally across the training pathway.

4.4.3 About 40% of the measures were established measures (i.e. data that is already in existence and tends to be available on a national level). The remaining measures would need to be collected at a local level using different data collection methods (including via a self-report approach). Other measures would be available through pre-validated questionnaires or scales, which can be customised with minimal development time.

4.4.4 All measures were potential measures of variability (differences between individuals, or in one individual over time) within the training pathway or in medical education. Approximately one quarter could be viewed as measures relating to variance in progression through training (i.e. exams) and the remainder as measures that relate to potential variance in experience or perception (i.e. increased awareness of cultural complexities). The latter type could be useful proximal measures\(^{15}\) of the impact of an intervention, whereas the former may be more distal measures\(^{16}\) in relation to intervention impact.

4.4.5 The majority of the measures were at the ‘micro’ or individual/group level, with few at the ‘meso’ or institutional level or the ‘macro’ or policy level. The expert stakeholder consultation was particularly useful in identifying organisational or system level measures (six additional measures were identified at this point). This reflects a recognition that differential attainment is the result of a variety of influencing factors, including the trainee, the trainer, the institution and the wider environment.

4.5 Finalisation of Measures

4.5.1 Via the stakeholder consultations and drawing on a process of ongoing review and validation by the research team, all measures were critically assessed against the criteria outlined in 4.3.4. Each measure’s overall suitability to indicate variations in medical education training pathways, and to help quantify the impact of interventions aimed at reducing variation and/or improving fairness, was also assessed.

4.5.2 Below is a summary of reasons for eliminating those measures not included as part of the final set of measures:

- A number of measures were combined (i.e. self-efficacy and confidence, social capital and social agency) based on similarity
- Some measures were deemed to overlap with other measures (e.g. a measure of ‘Optimal Programme Completion’ identified through the literature review was based on a US cohort and combined data on whether the student graduated, if they graduated in the minimal time, and their average A grade; all of which are separate measures already included)
- Some measures exhibited low relevance to indicating variations in medical education training pathways; i.e. a measure of ‘medical information utilization’ was a Nigerian

---

\(^{15}\) Proximal in this context means measures that can be collected immediately or very close after the end of an intervention

\(^{16}\) Distal in this context means measures that may only examine the impact of an intervention at some point in the future e.g. in a year’s time
based survey to explore how doctors search for and utilize medical information, which was deemed not suitable for inclusion.

4.5.3 The total number of measures in the final set of measures is 69. The measures are categorised based on a post-hoc categorisation process (i.e. broad themes identified after collation). Nine categories were defined and these are outlined in Figure 1 below. Six of the categories relate specifically to medical training whilst the remaining three categories relate to desired evaluation outcomes (learning and knowledge, behavioural and motivational and affect related measures based upon Kraiger, Ford and Salas’ (1993) evaluation model). Appendix C includes a list of the final 69 measures.

![Figure 1: Categorisation of Measures](image)

4.5.4 Below provides a summary of the categorisation of measures, including examples and key messages.

Selection Data

- This section contains **12 individual measures**, which are all quantitative, already established and nationally available, although some would reflect local variation (i.e. undergraduate interviews). These measures generally demonstrate good levels of reliability, validity and standardisation. Some evidence of differential attainment is observed in these measures.

- These measures cover a range of stages in the training pathway: school educational attainment (A-levels), selection into medical school (UKCAT), entry into Foundation (EPM, SJT), entry into Specialty (MRSA, interviews).

- Two measures (A-levels and UKCAT) were identified solely through the literature review and 10 were identified through interviews or other sources.

- **Key Message**: Predominately the measures identified in the selection category would be most appropriate for benchmarking rather than evaluation of an intervention, given their timing in the pathway (especially at an undergraduate level).
• **Key Message:** The measures are all at the ‘trainee’ level and, whilst still useful, may not be able to support broader interventions or programmes of work at the trainer or system level.

• **Key Message:** As all measures within this category are quantitative, there is a need to consider the minimum sample sizes required for each group being compared. Using samples of an appropriate size will enable robust analysis to take place and create confidence that any results are not due to chance. Statistical expertise should be sought to define the minimum sample size and supporting factors, if required.

• **Key Message:** If using selection data, it needs to be ensured that trainees do not think they are labelled as likely failures in future; ‘You are not [just] your MSRA score’ whilst still acknowledging that selection measures are still a valid approach.

**Examinations & Assessment**

• This section contains **5 individual measures**, which are all quantitative. Three are nationally available measures (ARCP, Work Place Based Assessment and Specialty Examinations), and these measures generally demonstrate good levels of reliability, validity and standardisation. Two further measures are available locally and are more variable in format and content (formative and summative assessments). Sub-categories within the examination measure were also identified, e.g. pass rates, number of re-sits and total score.

• Four of these were identified initially through the literature review and Work Place Based Assessment was identified through the interviews.

• The majority of the specific examinations identified took place at the postgraduate level; the broader assessments (e.g. summative) span the whole career pathway.

• **Key Message:** Specialty examinations are deemed to be most useful as a long-term evaluation measure because, while they are a measure of variation in progression, there is likely to be a time-lag between an intervention and access to examination results. This means that any evaluation could identify more embedded change as opposed to immediate, short-lasting reactions.

• **Key Message:** The measures are all at the ‘trainee’ level and, whilst still useful, may not be able to support broader interventions or programmes of work at the trainer or system level.

• **Key Message:** As all measures within this category are quantitative, there is a need to consider the minimum sample sizes required for each group being compared. Using samples of an appropriate size will enable robust analysis to take place and create confidence that any results are not due to chance. Statistical expertise should be sought to define the minimum sample size and supporting factors, if required.

• **Key Message:** If exam or assessment data is used, it needs to be ensured that trainees do not think they are labelled as likely failures in future based on an exam or assessment result.

**Training Progression**

• This section contains **9 individual measures**, which are all quantitative. These measures cover a range of training progression measures including completion of training stage, additional training time and access to additional learning support.
• The majority (7) are established measures available nationally and the existing data is expected to be reliable. Two of the measures (access to additional learning support, and formal remediation/disciplinary process) would need to be locally collected.

• Six of these measures were initially identified through the literature review and the remaining three identified through interviews or other sources.

• **Key Message:** This category contains some measures that tend to be at the end of the pathway (e.g. entry on the GMC specialist or GP register) and thus could only be used for long-term evaluation some years on from the intervention. Therefore, these measures may be less relevant in this context.

• **Key Message:** The measures are all at the ‘trainee’ level and, whilst still useful, may not be able to support broader interventions or programmes of work at the trainer or system level.

**Trainee Experience**

• This section contains 9 **individual measures** and cover a range of measures including trainee/trainer relationships, access to educational opportunities and relationship with institution.

• The majority of the measures would need to be collected locally, although some data is available through the National Training Survey and the Workforce Race Equality Standards (WRES) data.

• Eight of these were initially identified through the literature review, and 1 (frequency of trainer-trainee interactions) was identified through the interviews.

• **Key Message:** Some measures within this category could utilize a self-report methodology (individuals complete a scale or questionnaire rating themselves on a series of questions). Self-report measures have a number of advantages; in particular they may be viewed positively by the trainees as this type of assessment gives them a sense of ownership over their own assessment. For example, self-report measures may reduce feelings of stigmatisation, compared to sole use of an exam or selection score without additional contextual information. A limitation of self-report measures is the variable ability of an individual to accurately assess a change in their own knowledge and behaviour, or the potential for individuals to respond in a not entirely truthful manner due to fear or lack of trust.

**Trainer Experience**

• This section contains 6 **individual measures**, including general knowledge of DA in the deanery/LETB/trainer community and understanding how learning and training differs by country.

• All measures would need to be collected locally and could be collected using either a quantitative or qualitative format.

• One of the measures were initially identified through the literature review (trainers’ observations/perceptions) and the remaining five were identified through interview, expert stakeholder focus group or other sources.
• **Key Message**: The operationalisation of some of these measures is likely to depend on region and specialty.

**The Environment**

• This section contains 9 individual measures and include institutional climate, availability and/or quality of additional learning support, and leadership diversity.

• Only one was identified through the literature review, and the remaining eight were identified through interviews or other sources, including four from the expert stakeholder focus group.

• The majority of these measures would need to be collected locally, with the exception of some data possibly available within the WRES dataset (leadership diversity).

• **Key Message**: The operationalisation of this group of measures is likely to depend on region and specialty.

**Learning and Knowledge related measures**

• This section contains 7 individual measures, including knowledge level of the multiple factors believed to contribute to DA, and demystification of assessments. All measures need to be collected locally.

• All seven measures were initially identified through the literature review but expanded or developed within the interview process.

• **Key Message**: Whilst these measures will need to be collected locally, examples of sample questions and/or validated tools are available in the Measures section within this report. Pre-validated scales or questionnaires may make the evaluation design more robust and can also save time and effort.

**Behavioural related measures**

• This section contains 2 individual measures: proactivity and self-regulated learning. Both would need to be collected locally.

• One was identified through the literature review (self-regulated learning) and the other was identified through interviews.

• **Key Message**: Whilst these measures will need to be collected locally, examples of sample questions and/or validated tools are available and are provided in the Impact and Evaluation Guide. Pre-validated scales or questionnaires may make the evaluation design more robust and can also save time and effort.

**Motivational and Affect related measures**

• This section contains 10 individual measures including engagement with learning, resilience and burnout. The majority would need to be collected locally.

• All 10 measures were initially identified through the literature review.

• **Key Message**: Whilst these measures will need to be collected locally, examples of sample questions and/or validated tools are available and are provided in the Impact and Evaluation Guide. Pre-validated scales or questionnaires may make the evaluation design more robust and can also save time and effort.
• **Key Message:** Documenting perceived negative attributes could be viewed negatively by trainees, due to concerns about what may happen with the data in future. Careful consideration around engagement with these measures is likely to be required.
5. The Measures

5.1 Through the process outlined above, 69 measures have been included in the final set of measures. This means there is a great breadth of options, giving evaluators and others a variety of choice and flexibility when considering how to assess the impact of their intervention or programme of work.

5.2 It is acknowledged that there is limited definitive evidence in relation to potential measures, and a difference of opinion will exist in relation to the usefulness and suitability of the measures. All measures must be considered and interpreted in light of one’s own context and locality and thus used on a contextualised basis.

5.3 Whilst the research sought to identify a set of measures, the aim was also to critically assess the measures. This was focused on identifying information relating to the measure’s saliency and efficacy, so informed choices can be made about which measures would be most appropriate to employ. As such the following criteria have been used to further evaluate the measures.

A measure is most likely to be useful or appropriate if:

- If it is already available, it is easy to access, with minimal resource or time spent on gaining access including gaining ethics permissions
- If it is not already available, then it is easy to collect; i.e. a methodology is applied that is not time intensive and maximises engagement. Validated tools are also available that can be used without modification, or can be adapted which can save time and effort and make the evaluation design more robust
- If the data is already available, there is some evidence of its robustness including reliability and validity
- The data that is obtained is complete (i.e. not high levels of missing data), and is standardised, i.e. presented in a consistent, clear and ordered way (e.g. variables clearly identified and labelled)
- The measure is perceived favourably by trainees, trainers and other stakeholders. Some measures may be viewed less favourably; consideration of this at the outset and ways to overcome any reluctance regarding use must be considered. This may include active engagement with the participants about the purpose of the evaluation and reassurances about use of the data. Measures that may have more favourable perceptions could also be adopted. For example, self-report measures may reduce feelings of stigmatisation, compared to solely using an exam or selection score. Self-report may also be viewed positively by the trainees as this type of assessment gives them a sense of ownership over their own assessment
- Does not require extensive statistical expertise to interpret
- Is likely to be an accurate representation of what it aims to measure/assess i.e. not open to excessive manipulation, or dishonest self-reporting

5.4 The saliency, efficacy and overall appropriateness of a measure will also depend on the context in which it is to be used and there are a variety of influencing factors that will play a large part in
assessing the usefulness of a measure in any context. Therefore, when choosing which measures to use, the following should also be considered:

- **Intervention aim/outcomes:** the long-term aims of the intervention (i.e. what is the goal of the intervention, what does success look like in the short and long term) should inform the choice of measures. This includes who the intervention is aimed at/type of intervention i.e. trainee/trainer/the environment.

- At what point in the career pathway the intervention/programme of support is taking place and when is this timepoint relative to the measures that may be available. There are currently relatively limited measures that are directly related to medical school and Foundation training, although many of those identified could be applied universally across the training pathway. This is perhaps unsurprising as most evidence relating to DA has been found at the postgraduate level, given there are currently limited numbers of non-UK educated students in UK medical schools. However, an increase in non-UK undergraduate campuses may make any issue more prevalent or visible.

- **What is the available sample size?** The size of the intervention group will inform whether statistical analysis is possible. When dealing with quantitative data, there is a need to consider the minimum sample sizes required within each group being compared; this supports the completion of robust analysis and increases confidence results are not due to chance. Minimum sample sizes are dependent on a number of factors including required power\(^1\), statistical test used and the standard deviation\(^2\) of the variable of interest. Statistical expertise should be sought to define the minimum sample size and supporting factors, if required. If the numbers are very small, then indicative assumptions may be able to be made about the data, or the absence or presence of DA, but this would need to be qualified with qualitative measures/data before any conclusions could be made. When sample sizes are small, adopting a qualitative approach can be a good alternative.

- Whether **qualitative or quantitative** (or a combined approach) may be most beneficial. Qualitative approaches can add richness and depth to evaluation findings and are useful if only small sample sizes are available. However, they can be costly and resource intensive to implement but options are available to minimise resources required. Quantitative approaches are helpful in providing robust analysis and can be used to show statistically significant differences between groups. However, large sample sizes are needed, and statistical expertise may be required to support interpretation of findings. Combining qualitative and quantitative data can often be useful.

- If **multiple measures** would be appropriate to use. Each measure will have its own constraints in a particular context so, if resource is available, using multiple measures, metrics or data sources can help in mitigating any limitations, whilst also enabling triangulation of approaches to support conclusions and recommendations.

- Some measures **may not be available in particular regions or specialties**, or the way they are realised or attained may differ.

---

1. Statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error, or concluding there is no effect when, in fact, there is one, goes down
2. Standard deviation is a measure of the dispersion of a set of data from its mean
5.5 The measures are presented below classified by three categories based on their usefulness and/or appropriateness in relation to the points raised in 5.2 above, as well as the breadth or flexibility of the measure (i.e. how specific or generic the measure is with regards to what type of intervention it could be applied to). However, these are only broad categorisations and it may be that a measure that is deemed very useful in one context will not be useful in another context. Evaluators would be encouraged to review all measures in relation to their own specific context.

- **High** – likely to be the most useful or appropriate measures to use across a relatively broad range of interventions (n=15; Table 2)
- **Medium** – likely to have some usefulness across a range of interventions, but this may depend on influencing factors as to how useful or appropriate it is to use (n=33; Table 3)
- **Low** – may be of use in some specific contexts/interventions. Likely to have some limitations in application (n=21; Table 4)
**Table 2: Measures classified as High:** *likely to be the most useful or appropriate measures to use across a relatively broad range of interventions*

<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Accessibility &amp; Availability(^{19})</th>
<th>How robust(^{20}) is this as a measure?</th>
<th>What do I need to consider if using this measure(^{21})?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied, offered or accepted a place at GP or specialty</td>
<td>Selection</td>
<td>Good: GMC Progression Reports</td>
<td>Good: Accurate and standardised data</td>
<td>Interpretation consideration: Only uses Round 1 data</td>
</tr>
<tr>
<td>Multi-Specialty Recruitment Assessment (Clinical Problem Solving)</td>
<td>Selection</td>
<td>Medium: Held by the deanery and/or the specialty Royal College</td>
<td></td>
<td>Measurement consideration: This could be either number of attempts, pass rates or total score.</td>
</tr>
<tr>
<td>Multi-Specialty Recruitment Assessment (Situational Judgement Test)</td>
<td>Selection</td>
<td>Medium: Held by the deanery and/or the specialty Royal College</td>
<td>Good: Generally good evidence of this being a reliable and valid measure; national measures so will be standardised. Evidence of differential performance for some protected characteristic groups.</td>
<td>Measurement consideration: Consideration of which sitting should be used (first, last or averaged) should be undertaken at the outset. There is no correct answer; first sitting may be a more accurate reflection of true ability, whereas last sitting reflects final decision making. Engagement consideration: If using selection data, it needs to be ensured that trainees do not think they are labelled as likely failures in future; ‘You are not [just] your MSRA score’ whilst still acknowledging that selection measures are still a valid approach. Engagement consideration: There are some views that there may be bias within selection processes. Evidence generally shows that although there are differences, there are not inherent biases in the selection methods but perceptions of this bias are still likely to exist and may need to be managed.</td>
</tr>
</tbody>
</table>

\(^{19}\) Accessibility and availability is categorised into three classifications; Good – access and availability likely to be relatively straightforward and not too time intensive or complex to navigate, Medium – access and availability may be challenging, Low – access likely to be difficult and availability likely to be poor. Measures that are not in existence and will need to be collected will all be classified as Medium.

\(^{20}\) Robustness in this context generally refers to how reliable (provides consistent scores) and how valid (measures what it states to measure) the measure is, as well as reference to evidence of differential attainment. This also relates to how complete and standardised the data is.

\(^{21}\) This refers to a variety of considerations including perceived favourability by trainees/trainers/stakeholders, requirements for statistical expertise and whether it is likely to present an accurate representation of the measure.
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
</table>
| Specialty selection processes (i.e. interviews, OSCEs) | Selection              | Medium: Held by the deanery and/or the specialty Royal College | **Good:** Generally good evidence of this being a reliable and valid measure, but differences by specialty may vary. Evidence of differential performance for some protected characteristic groups. | **Engagement consideration:** If using selection data, it needs to be ensured that trainees do not think they are labelled as likely failures in future  
**Engagement consideration:** There are some views that there may be bias within selection processes. Evidence generally shows that although there are differences, there are not inherent biases in the selection methods but perceptions of this bias are still likely to exist and may need to be managed  |
| Combined selection scores e.g. MSRA and selection centre | Selection              | Medium: Held by the specialty Royal College and UKMED | **Good:** Available from a variety of sources, including the GMC progression reports and UKMED | **Interpretation consideration:** Different specialties will weight different elements differently |
| ARCP                                                | Examinations & Assessment | Good: Available from a variety of sources, including the GMC progression reports and UKMED | **Satisfactory:** Whilst ARCP requirements are consistent nationally, how they are applied may be impacted by local systems and processes and across specialties | **Measurement consideration:** Comprises of a number of elements, that you may wish to split out (i.e. exam data, WPBA data), which may be complex  
**Measurement consideration:** There can be limited variation within the scores, particularly at Foundation level where suboptimal scores are rare  
**Interpretation consideration:** ARCP outcomes may be influenced by the training received by assessors and the panel constituents. |
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility&lt;sup&gt;19&lt;/sup&gt;</th>
<th>How robust&lt;sup&gt;20&lt;/sup&gt; is this as a measure?</th>
<th>What do I need to consider if using this measure&lt;sup&gt;21&lt;/sup&gt;?</th>
</tr>
</thead>
</table>
| Specialty training exams; pass rates, number of re-sits and scores | Examinations & Assessment | **Good**: Pass rates available as part of GMC progression data; by PMQ awarding body, deanery and by demographic data Data available from Royal Colleges, although access at a local level may be variable. Available in UKMED | **Good**: Generally good evidence of reliability and validity. Evidence of differential performance for some protected characteristic groups. | **Design consideration**: Specialty examinations are deemed to be most useful as a long-term evaluation measure because, while they are a measure of variation in progression, there is likely to be a time-lag between an intervention and access to examination results. This means that any evaluation could identify more embedded change as opposed to immediate, short-lasting reactions.  
**Measurement consideration**: GMC progression reports do not currently show demographic data split by deanery so may have limited use for local evaluation.  
**Measurement consideration**: Statistical power is reduced for smaller programmes/regions.  
**Measurement consideration**: Rather than total score could use ‘score relative to pass mark’ i.e. how far above the pass mark they score, given that diets are not equated.  
**Measurement consideration**: Consideration of which sitting should be used (first, last or averaged) should be undertaken at the outset. There is no correct answer; first sitting may be a more accurate reflection of true ability, whereas last sitting reflects final decision making.  
**Measurement consideration**: If looking at number of re-sits, a measure such as ‘did they pass within x of years of training’ could be implemented, given that some take the exams first as a learning opportunity.  
**Interpretation consideration**: Whether candidates can self-nominate to take exams or not may impact pass rates. Do these ‘early takers’ who may use it as a gauge of their ability come from one specific demographic group?  
**Interpretation consideration**: Consideration of viability of using ‘total score’ when exams just require a minimum standard; though this view may vary dependent on specialty.  
**Engagement consideration**: There are some views that there may be bias within the examination processes; although there is no evidence that there are inherent biases, perceptions still exist that this is the case.  
**Engagement consideration**: If using exam or assessment data, it needs to be ensured that trainees do not think they are ‘labelled as likely failures’ in future, based on an exam or assessment result. |
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust(^{20}) is this as a measure?</th>
<th>What do I need to consider if using this measure(^{21})?</th>
</tr>
</thead>
</table>
| Length of extension/time to competence/less than full-time training            | Training Progression    | Low: At present, access to this data on a cohort basis is complex, but is available through ARCP/portfolio on an individual basis | Satisfactory: Data expected to be accurate | *Interpretation consideration:* Extension to training should not be viewed as a negative, and would need to be considered as part of a broader picture.  
*Interpretation consideration:* Should whether they have reached competence be a more important measure?  
*Interpretation consideration:* Reasons for the length of extension may be complex and not necessarily well documented. |
| Receipt of additional training time/number of extensions                       | Training Progression    | Good: Available through ARCP (outcome 3) | Satisfactory: Data expected to be reliable | *Interpretation consideration:* Reasons for extension may be complex and not necessarily well documented. |
| Trainee/trainer relationship                                                   | Trainee Experience      | Medium: May need to be locally collected. Some data related to this measure also available through the National Training Survey (i.e. *My educational supervisor is easily accessible should I need to contact them*) | Good: National Training Survey data comes from a standardised survey and will have been accurately analysed | *Measurement consideration:* Could include elements such as how supportive the relationship is, how encouraged they feel, degree of positive feedback received, degree of mentoring experienced, provision of timely and adequate feedback.  
*Measurement consideration:* Could be collected quantitatively through a survey, or through a qualitative method such as interviews.  
*Interpretation consideration:* May be difficult to influence the dynamic of such a relationship through an intervention.  
*Interpretation consideration:* Sometimes survey results are based on a small number of people; completion rates may be low for some regions where it is not mandatory.  
*Interpretation consideration:* Within survey data such as the NTS there is the potential for individuals to respond in a not entirely truthful manner due to fear of losing anonymity or lack of trust. |
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility&lt;sup&gt;19&lt;/sup&gt;</th>
<th>How robust&lt;sup&gt;20&lt;/sup&gt; is this as a measure?</th>
<th>What do I need to consider if using this measure&lt;sup&gt;21&lt;/sup&gt;?</th>
</tr>
</thead>
</table>
| Frequency of trainee/trainer interaction | Trainee Experience | **Medium**: Would need to be collected locally. Some data related to this measure also available through the National Training Survey (i.e. *The level of contact from my educational supervisor is appropriate for my training needs*) | **Satisfactory**: National Training Survey data comes from a standardised survey and will have been accurately analysed | *Interpretation consideration*: The frequency of interaction could be influenced by multiple factors isolating whether an intervention is having an impact could be challenging  
*Interpretation consideration*: Sometimes survey results are based on a small number of people; completion rates may be low for some regions where it is not mandatory  
*Interpretation consideration*: Within survey data such as the NTS there is the potential for individuals to respond in a not entirely truthful manner due to fear of losing anonymity or lack of trust |
| Positive attitude towards training/intervention | Motivation & Affect | **Medium**: Would need to be collected locally | **Unknown**: as data not known to be currently in existence | *Measurement consideration*: Would usually be collected as part of a post-evaluation survey to gauge reaction data |
| Engagement with learning | Motivation & Affect | **Medium**: Would need to be collected locally | **Unknown**: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid | *Measurement consideration*: There are multiple measures available for measuring student engagement. This reference provides a review of 11 methods for assessing engagement<sup>22</sup> |
| Resilience | Motivation & Affect | **Medium**: Would need to be collected locally | **Unknown**: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid | *Measurement consideration*: There are multiple measures available for measuring resilience. This reference provides a review of 17 resilience measures<sup>23</sup> |

<sup>22</sup> Fredricks, J and McColskey, W in S.L. Christenson et al. (eds.), Handbook of Research on Student Engagement,

<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidence/self-efficacy</td>
<td>Motivation &amp; Affect</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid</td>
<td>Measurement consideration: Could relate to a variety of aspects of training e.g. to continue life-long learning, in communication, in practice. Measurement consideration: There are a wide variety of confidence scales in the public domain. The most commonly used and validated self-efficacy scale is the Generalised Self-Efficacy Scale.</td>
</tr>
</tbody>
</table>

| A level (UCAS points)    | Selection         | Medium: Access through medical admission departments | Good: These tend to be viewed as reliable and valid assessments of ability. Evidence of differential performance for some protected characteristic groups. Potentially variability with A-levels taken outside of the UK, and across the different systems within the UK which may limit standardisation | Design consideration: This is relatively early on in the career pathway and therefore only likely to be useful for medical school interventions or programmes of work, or for benchmarking purposes |

Table 3: Measures classified as Medium; likely to have some usefulness across a range of interventions, but this may depend on influencing factors as to how useful or appropriate it is to use

---

<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Clinical Aptitude Test (UKCAT)/Bio-Medical Admissions Test (BMAT) scores</td>
<td>Selection</td>
<td><strong>Medium</strong>: Access through medical admission departments and available through UKMED(^{25})</td>
<td><strong>Good</strong>: Good evidence for reliability and validity. Evidence of differential performance for some protected characteristic groups.</td>
<td><em>Design consideration</em>: This is relatively early on in the career pathway and therefore only likely to be useful for medical school interventions or programmes of work, or for benchmarking purposes</td>
</tr>
<tr>
<td>Undergraduate interviews</td>
<td>Selection</td>
<td><strong>Medium</strong>: Access through medical admission departments and UKMED for some schools</td>
<td><strong>Satisfactory</strong>: Likely to be variable across schools</td>
<td><em>Design consideration</em>: This is relatively early on in the career pathway and therefore only likely to be useful for medical school interventions or programmes of work, or for benchmarking purposes</td>
</tr>
<tr>
<td>Foundation Programme Application System - Educational Performance Measure</td>
<td>Selection</td>
<td><strong>Medium</strong>: Available through UKFPO and UKMED</td>
<td><strong>Good</strong>: Generally good evidence of this being a reliable and valid measure; are national measures so will be standardised. Evidence of differential performance for some protected characteristic groups.</td>
<td><em>Design consideration</em>: This is relatively early on in the career pathway and therefore only likely to be useful for Foundation Programme interventions or programmes of work, or for benchmarking purposes</td>
</tr>
<tr>
<td>Foundation Programme Application System – Situational Judgement Test</td>
<td>Selection</td>
<td><strong>Medium</strong>: Available through UKFPO and UKMED</td>
<td></td>
<td><em>Engagement consideration</em>: There are some views that there may be bias within selection processes. Evidence generally shows that although there are differences, there are not inherent biases in the selection methods but perceptions of this bias are still likely to exist and may need to be managed</td>
</tr>
<tr>
<td>Foundation overall ranking</td>
<td>Selection</td>
<td><strong>Medium</strong>: Available through UKFPO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^{25}\) UKMED is the UK Medical Education Database and provides a platform for collating data on the performance of UK medical students and trainee doctors across their education and future career.
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
</table>
| Formative assessments | Exam & Assessment | Medium: Will be dependent on locality | Satisfactory: Will be variable dependent on the assessment | *Interpretation consideration:* Are mostly designed for learning and feedback thus are not held up to the same levels of scrutiny, evaluation or standard setting  
*Ethics consideration:* Permission/consent required to use in evaluation  
*Engagement consideration:* Consideration of perceptions of use recommended if using the data for something other than the original purpose i.e. assessments used for reflection purposes |
| Summative assessments | Exam & Assessment | Medium: Will be dependent on locality | Satisfactory: Likely to be good, though will likely be variable across institutions | *Ethics consideration:* Permission/consent required to use in evaluation |
| Work Place Based Assessments, including Multi-Source Feedback | Exam & Assessment | Medium: Accessed locally. Potentially through UKMED in the future | Satisfactory: Generally perceived as relevant and useful as an assessment tool; requirements are standardised although may be differences in how completed. May not be able to act as a true differentiator  
May be variability in credibility and accuracy of MSF | *Measurement consideration:* Is a broad measure, which will require thought as to which elements are used  
*Measurement consideration:* Data accessed individually making data collection resource intensive  
*Engagement consideration:* Careful and sensitive handling of these measures would be required. Need to consider trainees perceptions of using WPBA data for evaluation purposes, especially on an individual level i.e. feelings of stigmatisation |
| Repetition of year | Training Progression | Medium: Undergraduate measure; likely to be available through medical schools | Satisfactory: Data expected to be accurate | *Interpretation consideration:* Reasons for repetition of year may be complex and not necessarily well documented |
| Retention in training | Training Progression | Good: Available through ARCP (outcome 4) | Satisfactory: Data expected to be reliable | *Measurement consideration:* Numbers are likely to be small  
*Interpretation consideration:* Reasons for retention may be complex and not necessarily well documented |
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
</table>
| Access to additional learning support               | Training Progression            | **Medium**: Available at a local level from support unit or equivalent | **Satisfactory**: Data expected to be accurate       | **Interpretation consideration**: Additional learning support referral data is extremely variable across deaneries re: existence/budget/areas of coverage/quality, but where it does exist the richness of potential data is likely to be high  
**Interpretation consideration**: Both increase or decrease in referrals could be a measure of success depending on the intervention’s purpose |
<p>| Relationship with institutional environment; feeling valued and supported by organisation and NHS, including sense of belonging and inclusion. | Trainee Experience              | <strong>Medium</strong>: Would need to be locally collected | <strong>Unknown</strong>: as data not known to be currently in existence | <strong>Measurement consideration</strong>: Could be collected quantitatively through a survey, or through a qualitative method such as interviews. |
| Satisfaction or cultural connection with curricula   | Trainee Experience              | <strong>Medium</strong>: Would need to be locally collected | <strong>Unknown</strong>: as data not known to be currently in existence | <strong>Measurement consideration</strong>: Satisfaction with curricula could include user-friendliness of learning, teaching and assessment practice. Cultural connection with curricula relates to the role of an inclusive curriculum and involves not just the content of what is taught, but also questions of how it is designed, taught and assessed. Lack of cultural connection may stem from the design of curricula which may unconsciously marginalise specific groups |</p>
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
</table>
| Access to educational opportunities | Trainee Experience | **Medium**: Some data related to this measure is available through WRES data, NHS staff survey and the National Training Survey (i.e. questions relate to having time available to attend training, encouragement to take study leave and how gaps in the rota impact training opportunities) | **Good**: Survey data will have come from a standardised survey and will have been accurately analysed | Interpretation consideration: Sometimes survey results are based on a small number of people; completion rates may be low for some regions where it is not mandatory  
Interpretation consideration: Within survey data such as the NTS there is the potential for individuals to respond in a not entirely truthful manner due to fear of losing anonymity or lack of trust |
| Perceptions of the learning environment | Trainee Experience | **Medium**: Would need to be collected locally | **Unknown**: as data not known to be currently in existence | Measurement consideration: May include rating didactics, practical relevance, potential for professional development, and overall course structure and organization on 5-point Likert-type scales26 |
| Social capital and networks; *social factors can affect the learning experience of trainees and their engagement with learning* | Trainee Experience | **Medium**: Would need to be collected locally | **Unknown**: as data not known to be currently in existence | Measurement consideration: Social capital is multi-dimensional in nature and may include dimensions relating to degree of trust, social network structure and position, number of network memberships, association memberships and social participation, social connections and relationships, and the quantity or volume of social resources.  
Measurement consideration: At the individual level, social capital is usually measured by questionnaires27 |

---


<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trainers' observations/ perceptions (i.e. of trainees' confidence, if they are struggling, interaction with team members)</td>
<td>Trainer Experience</td>
<td><strong>Medium</strong>: Would need to be collected locally</td>
<td><strong>Unknown</strong>: as data not known to be currently in existence</td>
<td><em>Measurement consideration</em>: This will likely vary according to trainers’ skills and capability. <em>Measurement consideration</em>: Unlikely to be any formal measure or documentation of this.</td>
</tr>
<tr>
<td>Knowledge/ awareness of Differential Attainment and related issues in deanery/trainer community, and confidence in discussing</td>
<td>Trainer Experience</td>
<td><strong>Medium</strong>: Would need to be collected locally</td>
<td><strong>Unknown</strong>: as data not known to be currently in existence</td>
<td><em>Measurement consideration</em>: Typically collected through a questionnaire type approach or qualitative interview.</td>
</tr>
<tr>
<td>Trainer and educator understanding/ awareness of how learning and training/curricula differs by country</td>
<td>Trainer Experience</td>
<td><strong>Medium</strong>: Would need to be collected locally</td>
<td><strong>Unknown</strong>: as data not known to be currently in existence</td>
<td><em>Measurement consideration</em>: Could include knowledge and awareness of educational theory relating to learning.</td>
</tr>
<tr>
<td>Name of measure</td>
<td>Category</td>
<td>Availability &amp; Accessibility</td>
<td>How robust is this as a measure?</td>
<td>What do I need to consider if using this measure?</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------</td>
<td>-----------------------------------</td>
<td>------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Timely provision of adequate feedback from trainers i.e. as an outcome of</td>
<td>Trainer Experience</td>
<td>Medium: Would need to be</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Typically measured by questionnaire and contain a series of multiple</td>
</tr>
<tr>
<td>feedback training</td>
<td></td>
<td>collected locally</td>
<td></td>
<td>choice items grouped along one or more dimensions of the organisation. Items may relate to areas</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>such as innovation, satisfaction, leadership and communication</td>
</tr>
<tr>
<td>Institutional/educational climate, including perceptions and satisfaction</td>
<td>The Environment</td>
<td>Medium: Would need to be</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be</td>
<td>Measurement consideration: Validated measures exist, relating to organisational climate²⁸, as well</td>
</tr>
<tr>
<td></td>
<td></td>
<td>locally collected</td>
<td>standardised and likely to be reliable and valid</td>
<td>as learning/educational climate²⁹</td>
</tr>
<tr>
<td>Organisational understanding/awareness of Differential Attainment and related</td>
<td>The Environment</td>
<td>Medium: Would need to be</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Typically collected through a questionnaire type approach or qualitative</td>
</tr>
<tr>
<td>issues</td>
<td></td>
<td>locally collected</td>
<td></td>
<td>interview</td>
</tr>
<tr>
<td>Job satisfaction</td>
<td>Motivation &amp; Affect</td>
<td>Medium: Would need to be</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be</td>
<td>Measurement consideration: There are multiple measures available for job satisfaction. This</td>
</tr>
<tr>
<td></td>
<td></td>
<td>locally collected</td>
<td>standardised and likely to be reliable and valid</td>
<td>reference provides a review of job satisfaction measures and their reliability and validity to use</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>in hospital settings³⁰</td>
</tr>
</tbody>
</table>

²⁸ Patterson, M. et al. (2005) Validating the organizational climate measure: links to managerial practices, productivity and innovation. Journal of Organizational Behaviour, 26, 379-408
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>Motivation &amp; Affect</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid</td>
<td>Measurement consideration: There are a wide variety of stress scales in the public domain. The reference provides a summary and evidence related to different stress scales including perceived stress measures, daily event measures, and chronic stress measures. Engagement consideration: Documenting perceived negative attributes might be viewed negatively by trainees, due to concerns with what may happen with the data in the future.</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Motivation &amp; Affect</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid</td>
<td>Measurement consideration: Could be exam or general performance related Measurement consideration: Two validated screening tools commonly used are the Beck Anxiety Inventory (BAI) and the Generalised Anxiety Disorder (GAD-7) however these may not be appropriate in this context. Bespoke measures relating to the construct of interest may be more appropriate. Engagement consideration: Documenting perceived negative attributes might be viewed negatively by trainees, due to concerns with what may happen with the data in the future.</td>
</tr>
<tr>
<td>Burnout</td>
<td>Motivation &amp; Affect</td>
<td>Medium: Some questions as part of the National Training Survey (taken from the Copenhagen Burnout Inventory - CBI). Otherwise would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid</td>
<td>Measurement consideration: The two most common and validated tools are the Maslach Burnout Inventory (MBI) and the Copenhagen Burnout Inventory (CBI) Engagement consideration: Documenting perceived negative attributes might be viewed negatively by trainees, due to concerns with what may happen with the data in the future.</td>
</tr>
<tr>
<td>Communication competence/skills</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Could include elements such as use of open questions, tone, body language, colloquialisms</td>
</tr>
</tbody>
</table>

http://www.macses.ucsf.edu/research/psychosocial/stress.php
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding/awareness of cultural/societal differences in UK practice</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Could include areas relating to problem solving with patients, care of elderly relatives, disease focussed vs non-disease focussed, patient centred care</td>
</tr>
<tr>
<td>Clinical competence</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: This could be an objective measure, or trainee self-report</td>
</tr>
<tr>
<td>Trainee awareness of Differential Attainment and why there may be differences</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Typically collected through a questionnaire type approach or qualitative interview</td>
</tr>
<tr>
<td>Knowledge/understanding of assessments, including perceptions of bias</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: Typically collected through a questionnaire type approach or qualitative interview</td>
</tr>
<tr>
<td>General self-awareness; match between perceptions of competence and objective parameters</td>
<td>Knowledge &amp; Learning</td>
<td>Medium: Would need to be collected locally</td>
<td>Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid</td>
<td>Measurement consideration. There are a number of reviews that explore self-awareness and how to measure it 32 33</td>
</tr>
</tbody>
</table>

Table 4: Measures classified as Low: may be of use in some specific contexts/interventions. Likely to have some limitations in application.

---

<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red flags at selection</td>
<td>Selection</td>
<td>Low: May be available through specialty Royal College</td>
<td><strong>Satisfactory:</strong> Criteria for awarding a red flag is likely to differ by specialty, so may be variable</td>
<td><em>Measurement consideration:</em> The numbers are likely to be very limited and therefore this measure may be of limited use</td>
</tr>
<tr>
<td>Certificate of Completion of Training (CCT) or GP CCT completion</td>
<td>Training Progression</td>
<td><strong>Good:</strong> Available from Royal Colleges</td>
<td><strong>Satisfactory:</strong> Data expected to be accurate</td>
<td><em>Interpretation consideration:</em> This is at the end of training and potentially too far along the training pathway to be of use in this context</td>
</tr>
<tr>
<td>Entry on the GMC specialist or GP register</td>
<td>Training Progression</td>
<td><strong>Good:</strong> Available from the GMC</td>
<td><strong>Satisfactory:</strong> Data expected to be accurate</td>
<td><em>Interpretation consideration:</em> This is at the end of training and potentially too far along the training pathway to be of use in this context</td>
</tr>
<tr>
<td>Fitness to practise</td>
<td>Training Progression</td>
<td><strong>Good:</strong> Available from the GMC</td>
<td><strong>Satisfactory:</strong> Data expected to be accurate</td>
<td><em>Interpretation consideration:</em> This data is relatively sparse, so will have limited power</td>
</tr>
<tr>
<td>Formal remediation/disciplinary process</td>
<td>Training Progression</td>
<td><strong>Medium:</strong> Disciplinary process data <em>(Relative likelihood of BME staff entering the formal disciplinary process compared to white staff)</em> is available through Workforce Race Equality Standards (WRES) data. Other data would be locally collected</td>
<td><strong>Satisfactory:</strong> WRES data expected to be reliable</td>
<td><em>Measurement consideration:</em> The numbers are likely to be restricted and therefore this measure may be of limited use</td>
</tr>
<tr>
<td>Missed worktime/sickness reporting</td>
<td>Training Progression</td>
<td><strong>Medium:</strong> May be available at a local level</td>
<td><strong>Satisfactory:</strong> May be variable across institutions</td>
<td><em>Measurement consideration:</em> The numbers are likely to be restricted and therefore this measure may be of limited use <em>Interpretation consideration:</em> May wish to differentiate between chronic and long-term sick versus potential stress related sickness</td>
</tr>
<tr>
<td>Name of measure</td>
<td>Category</td>
<td>Availability &amp; Accessibility</td>
<td>How robust is this as a measure?</td>
<td>What do I need to consider if using this measure?</td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Experience of harassment, bullying or discrimination</td>
<td>Trainee Experience</td>
<td>Good: Some data related to this measure is available through WRES data, National Training Survey and NHS staff survey</td>
<td>Satisfactory: Data expected to be reliable</td>
<td>Measurement consideration: The numbers are likely to be restricted and therefore this measure may be of limited use</td>
</tr>
<tr>
<td>Allocation of trainee/trainer relationship</td>
<td>Trainer Experience</td>
<td>Low: May be available locally, but likely to need to be collected which may not be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: May relate to whether culture, gender and language are taking into account during the matching process Measurement consideration: The practicalities of obtaining and monitoring this data may be challenging</td>
</tr>
<tr>
<td>Training received by trainers</td>
<td>Trainer Experience</td>
<td>Low: Some aspects may be available locally, but if needs to be collected unlikely to be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: This could relate to trainer appraisals, what training has been received, and quality of training Measurement consideration: The operationalisation of this group of measures is likely to depend on region and speciality</td>
</tr>
<tr>
<td>Trainer demographics</td>
<td>Trainer Experience</td>
<td>Medium: May be available locally</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: May wish to explore following interventions related to allocation of trainers, or altering proportion of trainers from different demographic groups</td>
</tr>
<tr>
<td>Leadership diversity</td>
<td>The Environment</td>
<td>Medium: Some data related to this measure is available through WRES data (Percentage of staff in each of the AfC Bands 1 - 9 and VSM compared with the percentage of staff in the overall workforce)</td>
<td>Satisfactory: WRES data expected to be reliable</td>
<td>Measurement consideration: This would likely be a long-term measure and thus of limited use for immediate or medium-term impact evaluations</td>
</tr>
<tr>
<td>Individualised learning plans catered to the individual</td>
<td>The Environment</td>
<td>Low: May be available locally, but likely to need to be collected which may not be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td>Measurement consideration: This could be the existence of the learning plans, the number of learning plans or the quality of the learning plans. Note: currently unaware of this data being gathered or utilised in this way but has been deemed as a potential measure or type of measure</td>
</tr>
<tr>
<td>Name of measure</td>
<td>Category</td>
<td>Availability &amp; Accessibility</td>
<td>How robust is this as a measure?</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------</td>
<td>------------------------</td>
<td>------------------------------</td>
<td>---------------------------------</td>
<td></td>
</tr>
<tr>
<td>Additional learning support</td>
<td>The Environment</td>
<td>Low: May be available locally, but likely to need to be collected which may not be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td></td>
</tr>
<tr>
<td>Disciplinary procedures</td>
<td>The Environment</td>
<td>Low: May be available locally, but likely to need to be collected which may not be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td></td>
</tr>
<tr>
<td>Inductions</td>
<td>The Environment</td>
<td>Low: May be available locally, but likely to need to be collected which may not be straightforward</td>
<td>Unknown: as data not known to be currently in existence</td>
<td></td>
</tr>
<tr>
<td>Systems improvements; i.e. systems that better respond to the issues that are relevant for minority, ethnic and international graduates</td>
<td>The Environment</td>
<td>Medium: Would need to be locally collected</td>
<td>Unknown: as data not known to be currently in existence</td>
<td></td>
</tr>
</tbody>
</table>

**What do I need to consider if using this measure?**

*Interpretation consideration:* This systems level measure may relate to the availability or quality of support available.
Note: currently unaware of this data being gathered or utilised in this way but has been deemed as a potential measure or type of measure.

*Interpretation consideration:* This systems level measure may relate to the robustness and fairness of the disciplinary measures.
Note: currently unaware of this data being gathered or utilised in this way but has been deemed as a potential measure or type of measure.

*Interpretation consideration.* This systems level measure may relate to extent and quality of covering aspects relating to social and cultural differences in the induction programme, type and extent of social support offered as part of the programme and frequency of uptake of the programme.
Note: currently unaware of this data being gathered or utilised in this way but has been deemed as a potential measure or type of measure.

*Measurement consideration:* May wish to explore following system level interventions, and would be dependent on the intervention put in place.
<table>
<thead>
<tr>
<th>Name of measure</th>
<th>Category</th>
<th>Availability &amp; Accessibility</th>
<th>How robust is this as a measure?</th>
<th>What do I need to consider if using this measure?</th>
</tr>
</thead>
</table>
| Proactivity                            | Behavioural                     | Medium: Would need to be collected locally | Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid | Measurement consideration: Could relate to seeking help, learning or self-development  
Measurement consideration: An example of a validated instrument for measuring proactive behaviour is the Pro-active Personality Scale[^34] |
| Self-regulated learning                | Behavioural                     | Medium: Would need to be collected locally | Unknown: as data not known to be currently in existence, however validated tools will be standardised and likely to be reliable and valid | Measurement consideration: Could include self-generated thoughts, feelings, and actions for attaining academic goals  
Measurement consideration: There are multiple measures existing that examine different elements of self-regulated learning that could be drawn upon and/or adapted[^35] |
| Motivation for continued training      | Motivation & Affect             | Medium: Would need to be collected locally | Unknown: as data not known to be currently in existence | Measurement consideration: General motivation surveys exist, but likely that questions will need to be adapted to specific context |
| Future training aspirations            | Motivation & Affect             | Medium: Would need to be collected locally | Unknown: as data not known to be currently in existence | Measurement consideration: Could include for example, career goals, specialty choice, location (rural/urban)  
Measurement consideration: Typically collected through a questionnaire type approach or qualitative interview |
| Perceived preparedness                 | Learning & Knowledge           | Medium: Published by the GMC for Foundation Year 1 trainees only. Otherwise would need to be collected locally | Good: Data published by the GMC standardised and expected to be reliable | Measurement consideration: Drawn from an instrument devised by Fisher et al (2007[^36]) that measures trainees’ perceptions of competence, levels of knowledge and confidence related to course objectives.  
Measurement consideration: This is an example of how scales can be combined |

6. **Summary and Conclusions**

6.1 This research project sought to identify and critically evaluate a range of possible measures that may indicate variations in medical education training pathways and thus could be used for examining the impact of interventions aimed at reducing variation and/or improving fairness.

6.2 A multi-method approach to the research was conducted which incorporated a realist literature review, semi-structured interviews and focus groups with trainees, expert stakeholders and psychometricians. The multi-method approach increased the validity and reliability of the outputs by enabling the triangulation and validation of evidence gathered across the different methods. A total of 59\(^{37}\) individuals took part in the research.

6.3 The measures identified through the multi-method approach covered a breadth of outcomes, including variation in training outcomes or progression and variation in experiences and perceptions.

6.4 A total of 69 measures have been included in the final framework of measures, all of which were supported by multiple sources of data. The identification of 69 potential measures is very encouraging and suggests that there are a breadth of options available to employ in impact evaluation of an intervention or programme of work, providing flexibility and depth to an evaluation design. Having this breadth of measures will also enable greater opportunities for evaluation in the future.

6.5 The 69 measures were critically assessed in relation to their appropriateness and likely efficacy as evaluation measures. They have been presented grouped into three categories of likely usefulness or appropriateness, helping to guide an evaluator in their decision making whilst also taking their specific context into consideration.

6.6 This piece of work also sought to achieve objective 2; *To produce practical guidance to support organisations in evaluating interventions aimed at mitigating unfairness*. The Impact and Evaluation guide is available separately and seeks to incorporate relevant themes from this research process in addition to presentation of the measures, to help evaluators consider the broader context of any evaluative work.

6.7 The purpose of the guide is to support and facilitate the evaluation of interventions run in medical education with the long-term aim of expanding the evidence base in this area. An increased evidence base for medical education interventions designed to address the causes of DA will support learning from others and increase awareness of successful interventions or programmes of work. The guide is for anyone who is running or planning to run an intervention, initiative, training programme or programme of work, and wishes to understand if the intervention has had a meaningful impact.

6.8 The guide is divided into three sections:

- An introduction to the evaluation of interventions or programmes of work in the context of differential attainment

\(^{37}\) Two individuals took part in both the interviews, and the expert stakeholder focus group
- An overview of how to conduct an evaluation, with links to example material
- Examples of possible measures that could be utilised as part of an evaluation, including considerations to explore when using these.

6.9 The intention of the guide is that it is simple and easy to use, and that it provides access to further guidance or more detailed resources as required. The outlined approach to evaluation is guided by good practice principles but allows flexibility for individuals and institutions to define practical solutions that will work for them in their context.

6.10 The document is not intended to be a definitive guide to evaluation. Instead, it provides an overview of key points and theory concerning evaluation to support the robust evaluation of interventions and provides insight into a range of possible evaluation designs and measures.
Appendices

Appendix A: Detailed Overview of Methodology

Research Design

This research used a mixed-methods research design, combining a realist literature review with individual semi-structured interviews (various stakeholders), and focus groups (trainees, experts in differential attainment and psychometricians). The use of these combined methodologies increased the validity and reliability of the outputs by enabling the triangulation and validation of evidence gathered across the different methods. The realist literature review and the qualitative interview and focus group methodologies fed into both research objectives.

By using both individual interviews and focus groups for the qualitative data collection we combined the strengths of both data collection methods. Focus groups provide opportunities for rapid, diverse data collection that maximises interactions between participants and view sharing but also have the capacity for verification of conclusions and consensus drawing (Barbour, 2005; Cohen, Manion and Morrison, 2007). Whilst individual interviews have fewer opportunities for shared discussion, debate and spontaneity (Stringer, 2004), participants may be more candid expressing their views and perceptions concerning the potentially sensitive, personal and controversial nature of some of the issues relating to measuring differential attainment.

Thematic analysis was used to analyse the gathered data. Thematic analysis is a method for systematically identifying, organising and offering insight into patterns of meaning across a dataset, thus allowing the researcher to see and make sense of collective or shared meanings and experiences. The key stages when undertaking thematic analysis include; familiarising self with the data, generating initial codes, identifying emerging themes, verifying, confirming and qualifying the themes, defining and naming the themes and producing a report.

Realist Literature Review

The purpose of the literature review was the identification of possible measures of variability within the training pathway that may indicate DA (drawing on literature describing specific interventions, and medical education more broadly). A review of the organisational psychology literature was also undertaken to review models, measures and methods of evaluation. The purpose of the review was not to undertake an evaluation of the evidence related to any identified measures, but to identify and consider a range of potential measures for inclusion in the Measures Classification Framework.

This literature review employed a realist review approach; a systematic approach to analysing and making sense of existing research by considering the impact of context on observed outcomes. The benefit of a realist review is that it goes beyond simple description of data and explores the possible causal or environmental factors influencing outcomes, and that it supports the inclusion of published and unpublished documents and the “grey literature” (i.e., everything except peer-reviewed books and papers).

The review combined a review of the academic literature and information available from organisational documentation, to ensure that the information gleaned was both evidenced based

and focused on practical application. Academic sources included Medical Education journals, primarily UK based research but incorporating other comparable countries as required. Articles on selection, training and assessment were included, as was contemporary organisational psychology literature focused on the theory of training design and evaluation. Searches in the grey literature included stakeholder organisations and training/education provider websites, including broader educational websites, as well as policy documentation and conference presentations.

The search methodology involved three main stages: a) a search of databases, journals and additional search engines (e.g. Google Scholar) using the search terms. The web search included scanning websites of relevant key organisations for additional papers, articles and links to further sources. Citation tracking was carried out; checking reference lists for other relevant journal articles and following up references emerging from documents identified through the initial stages of the search; b) filtering by abstract; non-relevant papers were rejected at this stage if they were not relevant to the identified research aims. Following title screening, article abstracts were read and the article was kept for inclusion in the detailed review if relevant; c) a detailed review where the obtained papers were read by a member of the research team and a data extraction sheet was used to record relevant information such as method of identification (e.g. database), author, title, source (e.g. journal, conference), and key findings/points including potential measure.

**Interviews**

**Purpose:** The primary purpose of the interviews was to identify measures of variability and seek insight into their efficacy as a measure for the purposes of evaluation (objective 1). A secondary purpose was to explore the issues, challenges and barriers to evaluation of interventions/programmes of work, and considerations when implementing an evaluation within this context. The interviews also sought to explore the broader context of DA in relation to how it may also influence evaluation. This secondary purpose fed into objective 2.

**Approach:** Prior to commencing the interviews, an interview schedule was developed in conjunction with the GMC. This provided detailed overarching and probing questions to be used to guide the interviews and maximise the likelihood that valuable data was gleaned from each interview.

Telephone interviews were conducted to maximise convenience for participants and minimise travel costs. The interviews typically lasted for 45-60 minutes and were audio recorded (consent was sought at the commencement of the interview). Interviewees were provided with a briefing prior to the interview and assured anonymity in terms of their interview responses.

Four researchers undertook the interviews. On completion of the first five interviews, a debrief was held to review the evidence coming out of the interviews and calibrate the style and structure of the interviews. This preliminary analysis and discussion helped to shape further interviews using the same original questions, drawing upon a grounded theory methodology.

**Sampling:** The sampling strategy sought to target a range of individuals from different points in the career pathway, across different specialties and different regions in the UK. In practice, this was somewhat difficult to achieve, reflecting that DA has largely been identified at the postgraduate level, and that a significant proportion of the work in this area has been done within General Practice. Snowball sampling was used to generate a larger sample; respondents were asked for the

---

names of any further individuals who may be willing, and were in a position, to participate in the research.

A total of 41 one-to-one interviews took place and one virtual focus group consisting of five individuals (thus 46 individuals in total). Table 1 below provides an overview of the category, region, specialty and training pathway point of interviewees.

<table>
<thead>
<tr>
<th>Region</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>England</td>
<td>34</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td>2</td>
</tr>
<tr>
<td>Scotland</td>
<td>3</td>
</tr>
<tr>
<td>Wales</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specialty/Area</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>18</td>
</tr>
<tr>
<td>Physician</td>
<td>8</td>
</tr>
<tr>
<td>Anaesthetics</td>
<td>3</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>2</td>
</tr>
<tr>
<td>Emergency Medicine</td>
<td>1</td>
</tr>
<tr>
<td>Surgery</td>
<td>2</td>
</tr>
<tr>
<td>Paediatrics</td>
<td>1</td>
</tr>
<tr>
<td>Haematology</td>
<td>1</td>
</tr>
<tr>
<td>Other (E&amp;D/policy advisor, Training Support Unit)</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathway Point</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postgraduate</td>
<td>40</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1: Breakdown of interviewees

**Focus Groups**

Three formats of focus groups were held for groups of trainees, expert stakeholders, and psychometricians. The focus groups took place once the preliminary Measures Classification Framework had been as fully populated as possible using data from the interviews/literature review.

**Purpose:** Whilst the interviews supported data gathering and triangulation of findings from the literature review, the focus groups supported validation and critical analysis of the combined research outputs. The purpose of the trainee focus groups was to gain feedback from the trainees on the measures, particularly their perceptions of how appropriate they are to be used in this context and any considerations that should be taken into account. Trainee perceptions of ‘fairness’ of measures was an important aspect to consider when classifying the outputs from the literature review and interviews. This was because previous research has shown that trainee perceptions of fairness(bias) of measurement methods, particularly bias in recruitment and assessment, are critical to the issue of differential attainment, and may even be a causative factor underpinning differential attainment⁴⁰.

The purpose of the expert stakeholder focus group was to validate and confirm data and information collated from interviews and trainee focus groups. It was recognised that these earlier findings may be context specific, so ensuring consensus on the scope of evaluation measures of interest was an important step in ensuring the Impact and Evaluation Guide can be validated and agreed upon as suitable for use locally and nationally. Feedback was also sought in relation to the content of the Impact and Evaluation Guide. The psychometrician focus group was targeted specifically at discussing and classifying measures according to their psychometric properties.

**Sampling**: Volunteers for the trainee focus groups were sought through interview participants who were asked if they would be willing to either distribute an invite or host a focus group. Involvement of trainees who represented the protected characteristic groups under consideration, and who came from across the UK, was sought to ensure outputs were representative of these groups’ experiences in training. Two trainee focus groups were held; one in London and one in Newcastle. A total of 9 trainees took part in the focus groups. Volunteers for the expert stakeholder group were identified through the interview process, i.e. those individuals with a deep understanding of training pathways and the data available. Four experts took part in the expert stakeholder focus group. Two psychometricians took part in the psychometrician focus group. Thus, a total of 15 individuals took part on the focus group stage.

**Approach**: Focus groups were utilised as they provide opportunities for rapid, diverse data collection that maximises the interaction between participants and opportunity for view-sharing, but also have the capacity for verification of conclusions and consensus drawing. A schedule was devised for the trainee focus group, in addition to a briefing document covering the purpose of the session. Briefing material was also provided to the expert and psychometrician focus groups to facilitate discussions. The trainee focus groups were held face to face, while virtual focus groups were held for the expert stakeholder and psychometrician focus groups to maximise attendance.

**Qualitative Analysis**

On completion of audio transcription, all interviews and focus group outputs were coded. Thematic content analysis was employed for analysis of the interview and focus group data. Overarching themes, or structure codes, in relation to areas of interest as determined by the interview schedule were identified (i.e. general overview/understanding of differential attainment and how it is examined, examples of interventions and considerations, measures of variability) with additional layers of coding within these themes.

The initial framework was developed based on the first six interviews. Revisions to the initial coding framework took place iteratively throughout the capturing and review of data, to the point at which the research team was confident the coding structure was saturated (i.e. when all aspects of the data can be readily classified).

**Triangulation of Data**

**Objective 1**: As an output of the literature review, potentially relevant measures were captured in a preliminary Measures Classification Framework (see Results section for full details). This information was then triangulated with information emerging from the interviews and focus groups, which was used to update and progress the framework. Additional measures were either added into the framework or additional evidence about an already identified measure, including considerations for its use, were incorporated into the framework.
For some of the measures, evidence from the literature relating to the criteria was limited. Where this was the case, data was sought from further targeted research into these measures (i.e. internet searches relating to ARCP) and these areas were also focused on in the interview stage to help draw out more information and evidence as to their efficacy.

**Objective 2**: Themes and patterns in relation to broader areas relating to differential attainment, interventions and evaluation were identified and triangulated between interviewees, focus group attendees, and the literature review. These were then used to help inform the content of the Impact and Evaluation Guide.
Appendix B: Realist Literature Review Search Strategy & Scope

Sources

The literature review combined both a review of the academic literature and information available from organisational documentation, to ensure that information gathered was both evidenced based and focused on practical application.

Academic sources

• Medical Education journals, primarily UK based research but incorporated other comparable countries as required; included articles on selection, training and assessment

• Contemporary organisational psychology literature focused on the theory of training design and evaluation

• Peer-reviewed journals across other healthcare training programmes, including Nursing, Pharmacy, Dentistry, Midwifery

Conference presentations

• Recent presentations at AMEE, OTTAWA, CCME, IAMSE, ASPE etc., exploring interventions or investigations of differential attainment in medical training; primarily UK based research but incorporated other comparable countries as required

• Evidence from UK conferences designed to share best practice within medical training – DEMEC, Sharing Good Practice

Relevant stakeholder organisations and data sources

• GMC

• British Medical Association

• Networks such as BAPIO / BIDA / BSDO / MANSAG

• Medical Women’s Federation

Training providers

• Health Education England (HEE)

• 13 English LETBs

• NHS Education for Scotland

• Wales Deanery

• Northern Ireland Medical and Dental Training Agency (MDTA)

• NHS Employers

• Royal Colleges

• Academy of Medical Royal Colleges (AoMRC)

• UK Higher Education Funding Council for England (HEFCE)

Government
• Investigatory documents, policy documents, white papers on differential attainment in education

**Search Terms**

The review commenced with a thorough literature search strategy. Search terms/key words were identified by their relation to the Research Aims and included (“Training outcomes” OR “Differential attainment” OR “differential achievement” OR “academic achievement” OR “Protected characteristic*” OR “Fairness in training” OR “Medical Education” OR “Group difference*” OR “Ethnic* group performance”) AND (“Medic*” OR “healthcare”) AND (“practitioner” OR “doctor” OR “consultant” OR “Trainee” OR “medical student”)

The search terms used for the evaluation search included: “Evaluation” and “Training” or “Training outcome” or “Training transfer” or “Training Evaluation” or “Intervention Evaluation” and “Medic*”

A range of search terms for each of the identified search concepts was used, including:

- Synonyms, e.g. doctor, practitioner, consultant
- Acronyms, e.g. F1, Dr
- Differences in terminology across national boundaries
- Differences in spellings
- Old and new terminology
- Lay and medical terminology

Search terms were used individually initially and then in combination. Appropriate truncation for text-word searches will be used where applicable, e.g. nurs$ to find nurse, nurses, nursing etc.

**Scope: 5-10 years, prioritising more recent literature/data**

Research or findings from the last five years were prioritised. However, reflecting the finding that published studies relating to relevant interventions are limited, the search was expanded to incorporate research from the last ten years. During collation and assessment of measures, more recent literature was prioritised, recognising the definition of ‘differential attainment’ has evolved over the past few years.

**Database Search Strategies**

The review was carried out across databases including: PubMed, ResearchGate and Google Scholar, Medline, ERIC, CINAHL, Psycinfo, Embase and BioMed Central.

The search strategies created were designed for maximum sensitivity (recall) to ensure that any papers of significance were not overlooked.

The search uses a combination of text-words (free text) and subject heading searches. The search avoided restricting database subject heading searches using the ‘major descriptors’ or ‘subheading’ options in the first instance (to avoid missing relevant material). The strategy used specific functions in the relevant databases, for example the PubMed ‘related articles’ function and lists of references in CINAHL.
Advanced search options in databases were also used to increase sensitivity. This included using search operators to vary the combination of pre-determined search terms e.g. using AND to combine two different concepts; foundation doctor AND professional attributes, using OR to search for similar concepts; personal OR interpersonal qualities, using NOT to restrict search to particular concepts; foundation doctor NOT medical student etc.

Search Methodology

The search methodology used a structured ‘typology of evidence’ defined by Muir Gray⁴¹ to assess the literature identified and record its quality and relevance according to the type of study, using a number of predefined criteria such as effectiveness, salience and appropriateness. It involved three main stages; a) search, b) filtering by abstract, c) detailed review.

a) Databases were searched using the database search strategies outlined above. The web search includes scanning websites of relevant key organisations for additional papers, articles and links to further sources. Contact was made with expert individuals (through the interview process) to seek advice on unpublished or in progress work which was of relevance in order to obtain the most up to date and current material.

Citation tracking was carried out; checking reference lists for other relevant journal articles and following up references and contacts emerging from documents identified through the initial stages of the search.

b) All documents/articles were scanned for relevance on the basis of title and abstract; non-relevant papers were rejected at this stage. Strict criteria in relation to methodology and peer review were not considered paramount within this context, as the purpose of the literature review was to identify potential measures, rather than review, collate and evaluate evidence. In cases where an abstract was not available but the title suggests it could have been relevant, the full paper was obtained from electronic journals.

c) During the detailed review, the obtained papers were read by a member of the research team and data extraction sheets were used to record all relevant information. The data extraction sheets comprises of a table recording the following: method of identification (e.g. database), author, title, source (e.g. journal, conference), year, volume, pages, institutional affiliation, field, type of paper (e.g. empirical, conceptual or descriptive), research aim, study design, theory, analysis, key findings/points including relevant measure. If a paper was not felt to be relevant following this review, a note was captured to this effect. A second review of all relevant papers was then carried out by another member of the research team.

## Appendix C: List of Measures by Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Name of measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection Data</strong></td>
<td>A level (UCAS points)</td>
</tr>
<tr>
<td></td>
<td>UKCAT/BMAT scores</td>
</tr>
<tr>
<td></td>
<td>Undergraduate interviews</td>
</tr>
<tr>
<td></td>
<td>FPAS EPM</td>
</tr>
<tr>
<td></td>
<td>FPAS SJT</td>
</tr>
<tr>
<td></td>
<td>Foundation overall ranking</td>
</tr>
<tr>
<td></td>
<td>Applied, offered or accepted a place at GP or specialty</td>
</tr>
<tr>
<td></td>
<td>MSRA CPS</td>
</tr>
<tr>
<td></td>
<td>MSRA PD (SJT)</td>
</tr>
<tr>
<td></td>
<td>Specialty selection processes (i.e. interviews, OSCEs)</td>
</tr>
<tr>
<td></td>
<td>Combined selection scores e.g. MSRA and selection centre</td>
</tr>
<tr>
<td></td>
<td>Red flags at selection</td>
</tr>
<tr>
<td><strong>Examinations and Assessment</strong></td>
<td>ARCP</td>
</tr>
<tr>
<td></td>
<td>Formative assessments</td>
</tr>
<tr>
<td></td>
<td>Summative assessments</td>
</tr>
<tr>
<td></td>
<td>Work Place Based Assessments, including Multi-Source Feedback</td>
</tr>
<tr>
<td></td>
<td>Specialty training exams; pass rates, number of re-sits and scores</td>
</tr>
<tr>
<td><strong>Training Progression</strong></td>
<td>Length of extension/time to competence/less than full-time training</td>
</tr>
<tr>
<td></td>
<td>Receipt of additional training time/number of extensions</td>
</tr>
<tr>
<td></td>
<td>Repetition of year</td>
</tr>
<tr>
<td></td>
<td>Retention in training</td>
</tr>
<tr>
<td></td>
<td>CCT or GP CCT completion</td>
</tr>
<tr>
<td></td>
<td>Entry on the GMC specialist or GP register</td>
</tr>
<tr>
<td></td>
<td>Fitness to practise</td>
</tr>
<tr>
<td></td>
<td>Access to additional learning support</td>
</tr>
<tr>
<td></td>
<td>Formal remediation/disciplinary process</td>
</tr>
<tr>
<td><strong>Trainee Experience</strong></td>
<td>Trainee/trainer relationship</td>
</tr>
<tr>
<td></td>
<td>Frequency of trainee/trainer interaction</td>
</tr>
<tr>
<td></td>
<td>Relationship with institutional environment; feeling valued and supported by organisation and NHS, including sense of belonging and inclusion.</td>
</tr>
<tr>
<td></td>
<td>Satisfaction or cultural connection with curricula</td>
</tr>
<tr>
<td></td>
<td>Access to educational opportunities</td>
</tr>
<tr>
<td></td>
<td>Perceptions of the learning environment</td>
</tr>
<tr>
<td></td>
<td>Social capital and networks; <em>social factors can affect the learning</em></td>
</tr>
<tr>
<td>Trainer Experience</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Experience of trainees and their engagement with learning</td>
<td></td>
</tr>
<tr>
<td>Missed worktime/sickness reporting</td>
<td></td>
</tr>
<tr>
<td>Experience of harassment, bullying or discrimination</td>
<td></td>
</tr>
<tr>
<td>Trainers' observations/perceptions (i.e. of trainees' confidence, if they</td>
<td></td>
</tr>
<tr>
<td>are struggling, interaction with team members)</td>
<td></td>
</tr>
<tr>
<td>Knowledge/awareness of Differential Attainment and related issues in</td>
<td></td>
</tr>
<tr>
<td>deanery/trainer community, and confidence in discussing</td>
<td></td>
</tr>
<tr>
<td>Trainer and educator understanding/awareness of how learning and</td>
<td></td>
</tr>
<tr>
<td>training/curricula differs by country</td>
<td></td>
</tr>
<tr>
<td>Timely provision of adequate feedback from trainers i.e. as an outcome of</td>
<td></td>
</tr>
<tr>
<td>feedback training</td>
<td></td>
</tr>
<tr>
<td>Allocation of trainee/trainer relationship</td>
<td></td>
</tr>
<tr>
<td>Training received by trainers</td>
<td></td>
</tr>
<tr>
<td>The Environment</td>
<td></td>
</tr>
<tr>
<td>Institutional/educational climate, including perceptions and satisfaction</td>
<td></td>
</tr>
<tr>
<td>Organisational understanding/awareness of Differential Attainment and related</td>
<td></td>
</tr>
<tr>
<td>issues</td>
<td></td>
</tr>
<tr>
<td>Trainer demographics</td>
<td></td>
</tr>
<tr>
<td>Leadership diversity</td>
<td></td>
</tr>
<tr>
<td>Individualised learning plans catered to the individual</td>
<td></td>
</tr>
<tr>
<td>Additional learning support</td>
<td></td>
</tr>
<tr>
<td>Disciplinary procedures</td>
<td></td>
</tr>
<tr>
<td>Inductions</td>
<td></td>
</tr>
<tr>
<td>Systems improvements; i.e. systems that better respond to the issues that</td>
<td></td>
</tr>
<tr>
<td>are relevant for minority, ethnic and international graduates</td>
<td></td>
</tr>
<tr>
<td>Learning and Knowledge</td>
<td></td>
</tr>
<tr>
<td>Communication competence/skills</td>
<td></td>
</tr>
<tr>
<td>Understanding/awareness of cultural/societal differences in UK practice</td>
<td></td>
</tr>
<tr>
<td>Clinical competence</td>
<td></td>
</tr>
<tr>
<td>Trainee awareness of Differential Attainment and why there may be differences</td>
<td></td>
</tr>
<tr>
<td>Knowledge/understanding of assessments, including perceptions of bias</td>
<td></td>
</tr>
<tr>
<td>General self-awareness; match between perceptions of competence and objective</td>
<td></td>
</tr>
<tr>
<td>parameters</td>
<td></td>
</tr>
<tr>
<td>Perceived preparedness</td>
<td></td>
</tr>
<tr>
<td>Behavioural</td>
<td></td>
</tr>
<tr>
<td>Proactivity</td>
<td></td>
</tr>
<tr>
<td>Self-regulated learning</td>
<td></td>
</tr>
<tr>
<td>Motivational and Affect</td>
<td></td>
</tr>
<tr>
<td>Motivation for continued training</td>
<td></td>
</tr>
<tr>
<td>Future training aspirations</td>
<td></td>
</tr>
<tr>
<td>Positive attitude towards training/intervention</td>
<td></td>
</tr>
<tr>
<td>Engagement with learning</td>
<td></td>
</tr>
<tr>
<td>Resilience</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td></td>
</tr>
<tr>
<td>Job satisfaction</td>
<td></td>
</tr>
<tr>
<td>Confidence/self-efficacy</td>
<td></td>
</tr>
<tr>
<td>Stress</td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
</tr>
<tr>
<td>Burnout</td>
<td></td>
</tr>
</tbody>
</table>