Monitoring and evaluating impact

Equality Challenge Unit

Dr Amanda Aldercotte
Research Manager, ECU
info@ecu.ac.uk

To provide examples of evidence-based good practice and policy recommendations, practitioners need to evaluate the impact of equality and diversity initiatives. Unfortunately, limited resources in this area means formal evaluations of equality and diversity initiatives are rare, despite a growing need for such information.

Contents

This briefing aims to assist practitioners to design impact evaluations of their own equality and diversity initiatives first by clarifying what is meant by the term impact, and then by identifying both the traditional and alternative methods used to measure it. This briefing will also guide practitioners through the design process and provide insight into which methods to use, and when to use them.

As Equality Challenge Unit's (ECU) fifth research and data briefing, this briefing builds on our existing guidance on **working with equality data** by providing equality and diversity practitioners with an introduction to impact evaluation methodology and how to navigate this process.

Introduction

Over the past decade there has been a growing interest in impact evaluation in the realm of policy and, more recently, within the equality and diversity context as 'the evaluation gap' has become increasingly apparent (Gabarino and Holland, 2009). Evaluating the impact of an equality and diversity initiative is the only way to determine whether an **outcome** (eg a result or target) is truly related to an **intervention** (eg an initiative or action).

Impact evaluations also provide invaluable insight into whether equality and diversity initiatives can be improved upon (eg made more cost-efficient) or applied in other contexts (eg in other departments or institutions, or different equality areas and target groups). In essence, impact evaluations ensure better targeting, effectiveness and efficiency of initiatives to address inequalities.

However, evaluating impact in the context of equality and diversity is often impeded as the outcomes targeted by these initiatives are difficult to identify and quantify. For instance, it is problematic to assign numeric values to constructs such as staff members' understanding of promotion processes, or students' feelings of belonging. Further still, evaluating the impact of equality and diversity initiatives are particularly difficult because institutional change is both complex and a long-term process. This means practitioners need to decide which approach they are going to use to evaluate an initiative and also consider the additional factors that might influence their intended outcome and the appropriate timescale for allowing change to have taken place. Many of the methods used in other areas of social science and education research can be adapted for use in equality and diversity work.

This briefing begins with a brief set of key definitions followed by an overview of the quantitative methods traditionally used to measure impact. Next, it explores alternative methods that may be of particular use in equality and diversity work. It is important to note while these two sets of methodologies are presented separately, all of the evaluation approaches described in this briefing can be combined to complement one another. Finally, this briefing outlines the steps involved in designing an impact evaluation and provides additional resources for ongoing support.

Defining impact

One of the main hurdles in evaluating the impact of equality and diversity initiatives is the lack of clarity around what is meant by **impact**. In general, impact refers to an effect of an initiative, or **intervention**, whether it be positive or negative, direct or indirect, intended or unintended, on an outcome (Gabarino and Holland, 2009). At this point, it is also useful to distinguish between **output** (ie what is being produced such as programmes, training, or workshops) and an **outcome** (ie what is being achieved or produced by an output such as improved proportions of female staff applying for senior posts).

At times, the definition of impact is very broad and it can be difficult to pin down which outcome(s) should be the primary focus. For example, the Research Excellence Framework defines impact as 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, both in the UK or internationally, that goes beyond academia' (HEFCE, 2017). Other definitions of impact are narrower and context-specific, such as in Athena SWAN guidance documents where illustrating the impact of an equality and diversity initiative (or action) refers explicitly to gender equality outcomes and moving beyond simply reporting the initiative's **progress** (ECU, 2015).

Monitoring is the ongoing process of systematically collecting data on an outcome to check the initiative has been implemented correctly. Monitoring in itself is a means for measuring **progress**; in contrast, **evaluation** refers to the systematic assessment of an initiative, its design, implementation and results. While the two processes go hand-in-hand, as the information gathered during monitoring can be used in an evaluation, an evaluation should go beyond monitoring to describe an initiative's effectiveness (ie did it do what it intended to do?) and efficiency (ie did it do this well?) to assess its impact and sustainability.

Evaluating the impact of an intervention thus relies on some form of comparison, either before and after an intervention or with another group of individuals who were not exposed to an intervention (ie a control group). In technical terms, the 'before' assessment and the control group constitute the **counterfactual**, or what the target outcome would have been in the absence of the intervention. Comparing the counterfactual (or what would have happened to the outcome) to the factual (what actually happened to the outcome) is the key to impact evaluation. The next section of this briefing outlines the different types of counterfactual comparisons used in experimental research designs.

Figure 1: Are the comparisons in an impact evaluation the same as benchmarking?

Originally a business term, benchmarking is the process of comparing the performance of an individual or institution against another relevant group, institution or individual. Benchmarking typically includes a comparison with industry bests, or best practices from other companies. However, an organisation or institution can also benchmark against themselves by comparing their current performance to their past performance. This is a standard approach for many institutions, as seen in Athena SWAN applications for example. Regardless of whether institutions benchmark against each other or their past performance, these comparisons should be aspirational, identifying where there is room for improvement and setting ambitious goals.

While benchmarking can be used to identify targets, or, in the case of self-benchmarking, progress, this process does not constitute an impact evaluation in and of itself. This is because benchmarking, when used in isolation, does not require the same degree of scientific rigour as the traditional methods described in this briefing. Instead, benchmarking is one of many possible tools that can be applied in an impact evaluation as it can provide information on what targets an initiative should set and whether implemented initiatives are on track.

Traditional evaluation methods

There are two main sources of comparison from which impact can be assessed: (i) within a group of individuals over time, or (ii) between separate groups of individuals. To help illustrate how these two types of comparison would be applied to an equality and diversity initiative, an example of an initiative to improve staff awareness of promotion processes is presented alongside these methods of comparison (see figure 2).

It is worth noting that, in both types of comparison, the term **individual** can refer to individual people, departments or institutions, or even programmes or curricula. For instance, if an initiative is meant to diversify the curricula in a given department, the 'individuals' in this case would be the curricula of each course offered in that department and impact could be measured by monitoring indicators such as the number of black and minority ethnic (BME) or female references included in their course reading lists.

Figure 2: Improving the awareness of promotion processes among departmental staff

In this scenario, a department has identified that the number of female staff in early career posts (eg research assistants, teaching fellows, postdocs) applying for a promotion within the department is comparatively low to other departments in the higher education sector. After conducting a staff survey, the department finds female staff are not applying for promotion because they do not know enough about the department's promotion process (eg whether they can self-nominate themselves for a role or if they need to apply through their supervisor or line manager). To remedy this issue, the department rolls out a new advertising campaign to raise awareness on the different paths to promotion available.

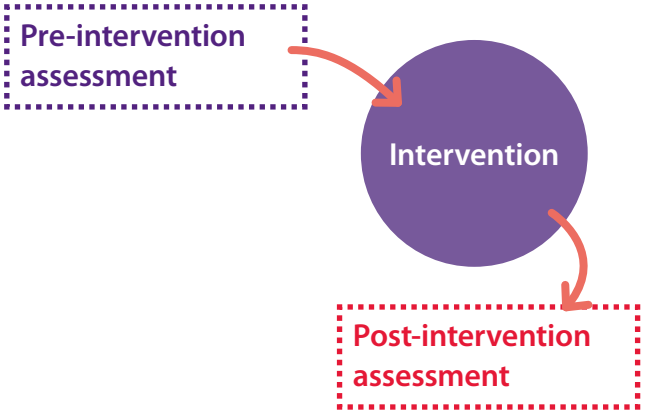
This campaign includes making the availability of a new post clearer via an all-staff email when a promotion round is taking place, as well as a series of posters displayed in the department. These posters not only provided additional resources for more information but also depicted success stories of senior members of staff who were willing to share their experiences, featuring mostly women. After another year of promotions, the department aims to evaluate whether this campaign had an impact on staff's awareness of promotion processes, particularly among female staff, and whether this in turn was related to their propensity to apply for promotion. How the traditional approaches to evaluation can be applied in this example are presented in figures 4 and 7.

Measuring differences over time within participants

A **within-groups design**, also known as a repeated-measures design, relies on measuring a target outcome both before and after an intervention. In this way, it is similar to the benchmarking against oneself using data from previous academic years that is **recommended by ECU**. Essentially, this approach looks at how participants in an intervention change over time using a pre- and a post-intervention assessment (figure 3).

Measuring differences over time requires **tracking** participants, or making sure an individual's responses in the 'before' assessment can be linked to their responses in the 'after' round of assessments. This requires collecting personal data, which has implications for data protection and storage, and specifically how participants' confidentiality and anonymity will be maintained.

Figure 3: Within-groups (repeated-measures) design



One common approach is to anonymise the survey responses by replacing participants' personal details with a unique identification number (also known as **pseudonymisation**) (see figure 4). A master file matching participants to their ID numbers will need to be stored in a secure place that is separate from the survey responses. Relabelling participants' surveys with a set of ID numbers will allow their 'before' and 'after' responses to be linked without risking participants being personally identified.

A repeated-measures ANalysis Of VAriance (ANOVA), also known as a within-subjects ANOVA, is the main quantitative approach to analysing within-groups data as it tests whether there is a statistically significant difference in participants' scores from the pre- and post-intervention assessments. In this type of analysis, factors that are expected to have an influence on an outcome (eg being exposed to an intervention) are considered an independent variable, or predictor, while other factors that might be related to the outcome, such as working part- versus full-time, are defined as control variables. To do this, data need to be entered into a database in a certain way; specifically, each row of data should refer to a single participant with individual columns representing pre- and post-intervention assessments (see figure 5). It is also worth noting that this type of analysis requires a continuous outcome measure, where any value is possible and not limited to two or three categories. In other words, a Likert scale that asks participants to rate their awareness of promotion processes on a scale of 1 through 7 would be acceptable but a 'yes' or 'no' response would not.

Figure 4: Improved staff awareness over time.

To determine whether the advertising campaign improved staff members' awareness of promotion processes, and in particular whether the impact of this campaign differed for male and female staff, the department conducts two rounds of another survey to gauge staff members' understanding in greater detail. Staff are asked to complete this survey before the campaign is rolled-out, as well as after the campaign has run its course.

This survey includes questions about staff's degree of understanding, whether they know about specific promotion processes, and other questions around how likely they are to apply for a promotion in the next round. Also included in this survey are staff members' names, contact information and personal details (such as their contract mode and type); this is so the department can link individual participant's pre- and post-intervention survey responses to each other.

The pre- and post-intervention surveys are identical except the post-survey also includes questions directly related to the intervention (eg 'Did you receive and open an email from the department head notifying staff of the last promotion round?' and 'The department has put a series of posters depicting senior staff member's promotion success stories. To what degree did you notice these?'

To analyse these data, the department inputs the results of their two surveys into a dataset (see figure 5). Data could be entered into a spreadsheet in Microsoft Excel, but to conduct a repeated-measures ANOVA a more sophisticated statistical software package, such as IBM's ***SPSS*** or the free, open source program ***R***, must be used (see figure 9 for resources related to these packages). Using one of these software packages, the department runs a repeated-measures ANOVA to determine whether staff members' exposure to the all-staff email and campaign posters is related to their post-intervention awareness of promotion processes. The key to a repeated-measures ANOVA allows participants' pre-intervention awareness to be taken into account. Essentially, this approach answers the question: did the campaign improve staff members' awareness of our department's promotion processes over and above what staff already knew about these processes?

A repeated-measures ANOVA would also tell the department whether the campaign had a different impact on male and female staff members' awareness by adding participant sex as an independent variable.

Figure 5: Example of how to input data for a within-groups analysis

ID	Sex 0 = Male 1 = Female	Pre-Q1 1 = SD 7 = SA	Pre-Q2 1 = SD 7 = SA	Post-Q1 1 = SD 7 = SA	Post-Q2 1 = SD 7 = SA	Intervention A 0 = no 1 = yes	Intervention B 0 = did not see posters 1 = noticed but did not read 2 = read posters
0010	0	2	4	5	6	0	2
0011	1	4	5	6	5	1	0
0012	0	3	4	5	5	1	1

SD: strongly disagree, SA: strongly agree

Intervention A: opened and read promotion email
Intervention B: exposure to campaign posters

There are two major advantages to measuring differences over time within the same sample. First, employing a within-groups design is typically more cost-effective than other experimental designs as this approach does not require recruiting additional participants for a control group. By comparing participants to themselves, the number of people needed in the evaluation is smaller, which will be financially beneficial as well as save on time. Second, using participants' pre-intervention responses as a benchmark reduces the amount of error or 'noise' in the analysis. Participants come from a variety of backgrounds and have their own unique experiences that will influence how they respond to questions asked. However, when a target outcome is measured twice within the same individual, the researcher takes the influence of background factors (eg socioeconomic background, personal experiences) into account by keeping them constant over both rounds of assessment.

The main disadvantage of a within-groups design is this approach cannot be used to determine **causality** (the change in the outcome is caused by the intervention). In our example, we cannot say our advertising campaign caused greater staff awareness of promotion processes as we cannot rule out the possibility that staff would have had improved awareness even in the absence of our campaign.

Moreover, while within-groups designs help reduce the amount of noise in the analysis it is still important to consider the other factors that could have an influence on the targeted outcome and take these into account. For instance, the sample of participants might include staff on part-time as well as full-time contracts, meaning that not all participants will have the same degree of exposure to the on-site advertisements. As such, it would be useful to statistically take this influence into account when assessing whether or not the advertising campaign has increased staff awareness of promotion processes. This is one reason why repeated-measures ANOVA is the main approach to analysing within-groups data as it allows for the statistical control of other factors related to the targeted outcome.

Comparing participants to a control group

But what if it is not possible to track the same sample of staff over time? For instance, if the advertising campaign for promotions processes was implemented across an entire department and we could not be assured that the same staff members would complete both surveys?

If it is not possible to track the same participants over time, an alternative is a **between-groups design** which measures differences in an intervention group relative to a control (or comparison, or experimental) group. What makes this approach experimental is one group is exposed to an intervention (or special treatment) to be compared with another group that has been kept constant.

If participants are randomly assigned to either the control group or to the intervention group, this becomes a **randomised control trial** from which causal links between an intervention and outcome can be inferred. Random assignment to either group eliminates the systematic differences in the known (and unknown) characteristics that could influence the target outcome. In other words, if participants are randomly assigned to one of the two groups, both groups have the same odds of including participants with the different characteristics that might be related to the outcome, such as the type of contract they hold, their family background, whether they have children, and so on. Having equal probabilities across the two groups essentially cancels out the influence of other factors on the outcome, because both groups are equally likely to have this noise in their analysis.

Although randomised control trials are more difficult to employ in social sciences compared to medical or physical sciences, it is definitely possible to apply this method of evaluation in an equality and diversity context. For example, a department wants to evaluate whether a mentorship scheme for postdocs improved their awareness of career development opportunities and makes them more confident to apply for lectureship posts. In this example, it is possible to randomly assign half of the postdocs who applied to the scheme to the intervention group (ie they are assigned a trained mentor), with the other half being assigned to the control group. This approach would allow the department to compare the progress of the postdocs who were assigned a mentor to those who were not assigned a mentor. However, in this example, it may be more appropriate to use a **waitlist control group**, as it would be questionable to withhold this possible benefit from students on an ethical level. The waitlist control group of postdocs would be assigned a mentor, but only after the comparison with the intervention group was complete.

A between-subjects ANOVA (also referred to as one-way independent ANOVA or factorial ANOVA) is typically used to explore group differences in between-groups data. This approach determines whether two or more groups (or individuals) differ in a target outcome. Applying this method to the example of postdocs in a mentorship scheme would determine whether postdocs who were assigned a mentor were significantly more likely than those without a mentor to feel that they understood their career development opportunities and were more confident to apply for lectureship posts. Similar to the repeated-measures ANOVA used for within-groups analyses, between-subjects ANOVAs allow extraneous factors that might be related to the target outcome (such as subject area or year of study) to be statistically taken into account. These variables are labelled as control variables, while the factors that are expected to impact an outcome being tested are labelled as independent variables. Although the labels in a between-subjects ANOVA are the same as the labels for variables in a repeated-measures ANOVA, how data are inputted for a between-subjects ANOVA is different: the target outcome is presented in a single column (instead of a pair of pre- and post-intervention columns) while a second column summarises to which group (intervention or control) each participant belongs (see figure 6). However, each row still represents a single, anonymised participant.

All types of ANOVA require the outcome variable to be on a continuous scale. When the outcome measure is limited to one or two categories (ie a categorical variable, such as either applying for promotion or not applying) it is more useful to use alternative quantitative methods such as logistic regression (see ECU's research and data briefing on **intersectionality** for a description of this type of analysis).

Figure 6: Example of how to input data for a between-groups analysis.

ID	Sex 0 = Male 1 = Female	Group 0 = control 1 = intervention	Q1 1 = SD 7 = SA	Q2 1 = SD 7 = SA
0010	0	1	2	5
0011	1	0	5	7
0012	0	0	4	4
0013	1	1	3	4

SD: strongly disagree, SA: strongly agree

While randomised control trials are considered the gold standard for determining causality, they are not always practical in social science research. This is the main disadvantage of randomised control trials, as their applicability is limited by the type of sample required and how an intervention is implemented. For instance, in our evaluation of the promotion processes advertising campaign, it would not be feasible to randomly assign staff to the intervention (exposed to the ads) or control groups (not exposed to the ads).

Instead, a more appropriate approach would be to make use of naturally occurring control groups that are convenient and accessible, such as comparing staff exposed to the advertising campaign to another department that does not have this initiative in place (see figure 7).

Figure 7: Improved staff awareness compared with another department.

To examine whether the advertising campaign improved staff members' awareness of promotion processes, the department decides to compare their staff members (ie the intervention group) to another sample of staff in a similar department at their institution that does not advertise its promotion processes (ie the control group). Again, a survey including questions about staff's degree of understanding, whether they know about specific promotion processes, and how likely they are to apply for a promotion in the next round is sent to both sets of staff.

Direct comparability between the responses from the intervention group and the control group is key: both sets of staff completed the same survey. Also included in this survey were additional details on participant's contracts (eg full- versus part-time, current grade, etc), how long they have been in the department/institution, and other demographic information (eg age and gender). Staff in both departments complete the survey after the campaign has run its course.

Once the survey is complete, the department runs a between-subjects ANOVA using a statistical software package such as ***SPSS*** or ***R*** (see figure 9 for resources related to these packages). To which group staff members belong is entered as an independent variable, while staff responses to the survey items around awareness and understanding are entered as dependent variables. The department also includes details on staff members' contracts and length of service as control variables. This analysis then answers the question: After rolling out a new advertising campaign, were staff in our department more aware of how to apply for promotion than staff in another department, regardless of how much time they spend on site

and how much time they have been working at the department?

To identify whether male and female staff experienced the campaign differently, the department would include information on participant's sex as an independent variable in their analysis. However, to determine whether any differences between male and female participants' awareness are specific to their own department, they would need to create what is called an **interaction term** to distinguish between male and female staff in their own department (the intervention group) and the department they are comparing themselves to (the control group). An interaction term is essentially the product of the two independent variables (in this case, the 'Participant sex' variable multiplied by our 'Group' variable in figure 6). See ECU's research and data briefing on **intersectionality** for more information on how interaction terms can be used in equality and diversity evaluations.

What if the department could not survey another department's staff?

If the department did not have access to another department's staff, it would have to base its evaluation on a different outcome measure, such as comparing how many members of staff apply for promotion to a previous academic year or perhaps another department who make this information public. Reliance on this kind of data would make the evaluation method more of a benchmarking exercise instead of a between-groups analysis. The difference lies in the degree of rigour employed in these two approaches; comparing participants' responses on a survey allows more sophisticated quantitative methods to be used (that can control other factors of influence), while benchmarking does not allow for this, which makes it difficult to directly link the initiative (ie an advertising campaign) to the outcome (ie improved awareness). An important component to this type of evaluation would then be including another, complementary method that would facilitate linking the initiative to the outcome, such as conducting interviews or focus groups.

Of the three types of comparisons described in this briefing (within-groups, and the two types of between-groups comparisons described in the next section), the only method from which causality can be inferred is a randomised control trial.

A between-groups design in which participants are not randomly assigned to the control group cannot be used to determine causality because it is impossible to rule out whether other, unmeasured factors might have produced the change in the target outcome. Nonetheless, this methodology is widely adopted in the social sciences because there are a number of ways researchers and practitioners can strengthen their results.

The first way to improve the rigour of using a convenient control group is to make sure that participants in the control group are as similar as possible to those in the intervention group. For instance, when looking at equality and diversity initiatives within an institution or department, it is useful to consider matching participants based on certain characteristics, such as: the relative size of the institution or department; the age, gender and ethnic composition of staff; specific subject areas; and staff contracts (eg fixed-term or open-ended, part- or full-time, early career or senior management). Taking time to match the intervention and control groups as closely as possible reduces the amount of variability in the outcome variable influenced by outside factors and limits the amount of noise in subsequent analysis. Thus, the use of randomised control trials is possible in equality and diversity research

but it depends on the sample and format of the initiative involved. Similarly, researchers can improve the strength of their results by measuring the different characteristics and factors (such as those listed above) that might have an influence on the outcome. If these are measured, they can be statistically taken into account in a between-groups ANOVA (described above).

Finally, the third way researchers can strengthen a between-groups analysis that does not include a randomly assigned control group is to combine this approach with a within-groups analysis. This approach is particularly useful when it would be unethical to withhold an intervention from participants. In the example above evaluating the effectiveness of a mentorship programme for postgraduate students, both the intervention group and control group would complete pre- and post-intervention assessments as they would in a between-groups design, but the control group would be assigned mentors after the post-intervention assessment and be assessed a final time after completing the programme (see figure 8).

Figure 8: Combining within- and between-groups methods.

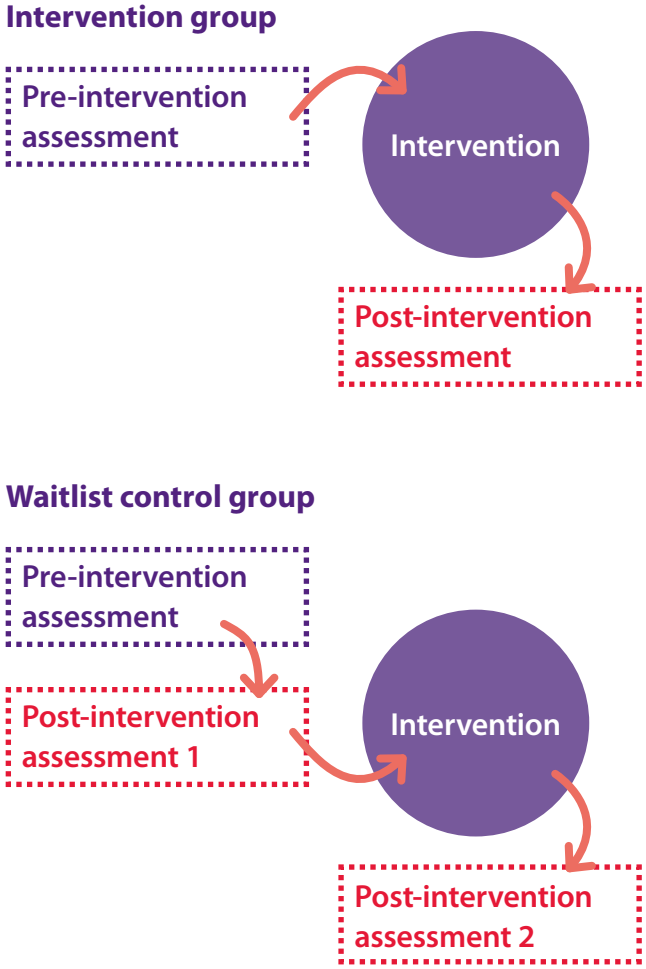


Figure 9: Resources and guides on how to conduct repeated-measures and within-subjects ANOVAs in *SPSS* and *R*.

***Social Research Methods* by Alan Bryman (5th edition)**
This textbook includes self-contained chapters explaining different evaluation methods, their individual pros and cons, as well as how to get started with using them.

Discovering statistics
This website brings together a number of free resources that have corresponding textbooks which provide more detailed instructions on how to conduct quantitative analyses using statistical software, such as ***SPSS*** (Field, 2009) and ***R*** (Field, Miles and Field, 2012). These textbooks also include accessible information on the maths behind these techniques. This website includes a particularly useful **guide to conducting repeated-measures ANOVA**.

Alternative approaches

The following section outlines alternative methods to the more traditional forms of impact evaluation described above. As the overarching aim of equality and diversity work is to address structural barriers, the majority of equality and diversity initiatives are smaller in scale and aim to bring about change within a specific group of people, context or institution. When working with smaller samples or exploring the impact of a new initiative, these alternative evaluation methods may not only be more appropriate from a statistical point-of-view but also more informative about an initiative's effectiveness and efficiency.

Outcome mapping

Outcome mapping is a way of assessing projects that are trying to bring about 'tangible' change. Outcome mapping is not a stand-alone methodology, instead it is a method for **planning** an evaluation as it provides a set of tools that can be combined with other quantitative or qualitative methods.

Outcome mapping helps researchers be specific about who the target participants are, what changes the initiative hopes to bring about and which strategies should be used to do so. In this approach, the individuals, groups and organisations that interact directly with an initiative are referred to as **boundary partners**; these people are, in other words, the people who we expect an initiative to influence. The 'outcome' is referred to as **behavioural change**, which can encompass changes in the behaviour, relationships, activities or actions of the people working directly with an initiative (Earl et al., 2001).

Although outcome mapping can be used to link behavioural changes to an initiative, this methodology does not allow causality to be inferred. Instead, this process identifies an initiative's **contributions** to an outcome. This is achieved by setting a target outcome, referred to as a **challenge statement** (what the final desired outcome is meant to be or should look like), as well as **progress markers** (a set of statements describing how the outcome is expected to change gradually until the target is attained).

Outcome mapping asks project planners to answer four questions:

- why are we doing this initiative?
- who are the boundary partners?
- what are the tangible changes we want to bring about?
- how will this initiative bring about behavioural change among the boundary partners?

While outcome mapping does not specify a particular method for measuring behavioural change, qualitative data is typically collected through the use of self-assessments, referred to as **journals**. These journals log information about progress markers, the activities completed and the internal performance of the initiative.

Outcome mapping is a cyclical process as it uses issues identified in the journals to inform the next iteration of the intervention. As such, it is a valuable method for evaluating new initiatives or initiatives that target a smaller group of individuals so the data collected can be used to inform, refine and improve. However, it is important to note that outcome mapping involves continuous monitoring of behavioural change, which requires a larger amount of time and resources. Thus, successfully using this type of evaluation method requires support from higher levels of an organisation and a commitment to investing in tangible change.

Outcome mapping is unique as it embeds how an outcome will be monitored in the initiative, and focuses on collecting data on immediate, basic changes that lead to longer-term, transformative change.

Case studies

Case studies focus intensively on a particular 'case', be it an individual, group of individuals or institution/department, examining this unit as a distinct whole. Rich, detailed information about a participant or place is collected. Case studies can employ a variety of quantitative and qualitative methods, ranging from structured questionnaires, semi-structured interviews, completely unstructured interviews, observation to document analysis (figure 10). This information is used to answer the 'what happened?' question, as it allows the person conducting the study 'to create a full, complex picture of what occurred' (Balbach, 1999).

Figure 10: Methods of collecting data.
Although these methods are described in terms of their utility within a case study evaluation, each type of assessment can be used in the other, traditional and alternative, methods outlined in this briefing. As many of these involve discussions with individuals or groups of participants, these methods are typically audio- or video-recorded for later analysis.

Structured questionnaires. Creating a highly structured questionnaire requires the researcher to know all of the pertinent questions and to have thought of all of their possible responses. As such, highly structured surveys are not typically used in case studies. Instead, questionnaires can be used to obtain basic information (eg demographic information about participants or degree of exposure to an initiative's actions) that would complement data obtained through other methods and allow some degree of comparison across participants.

Semi-structured interviews. Probably the most common approach to collecting data in a case study, semi-structured interviews specify a list of general topics but include open-ended questions (meaning that the researcher does not need to predict all possible responses beforehand). Specifying a list of topics to be covered in the interview ensures that similar information is collected from all participants, which in turn allows researchers to compare these responses to one another. As it is semi-structured, it is not necessary for all of the topics to be discussed in the same order or asked using the same wording.

Unstructured interviews. As the name suggests, unstructured interviews remove the skeleton structure applied in semi-structured interviews, allowing for maximum flexibility in the experiences discussed. This type of interview is most useful as a way of obtaining initial information that will be used to inform more in-depth and targeted data collection, for instance when a researcher knows little about the participants targeted by an initiative or the context in which that initiative will be implemented.

Focus groups. This method of data collection uses a small group of usually six to 10 people to gauge their opinions or hear their experiences of a particular issue or theme. The format of a focus group is to encourage participants to engage and shape each other's thoughts, forming opinions as a collective or clarifying individual differences through their disagreements. Focus groups typically use a list of questions, similar to that used in a semi-structured interview, as a discussion guide for the group conversation. However, focus group discussions can also be unstructured, as the main purpose of this approach is to allow interaction among participants with minimal interference from the facilitator.

Observation. Sometimes referred to as **naturalistic observation**, this approach includes the evaluator putting themselves in the context or organisation that they are trying

to evaluate. For instance, in our example of trying to raise staff awareness of promotion processes through an advertising campaign, an evaluator might place themselves in the public spaces where the advertisements are being displayed and observe staff responses, recording this information in their field notes. In this way, observational approaches allow the evaluator to directly experience the context they are trying to investigate.

Participant observation. In contrast to the passive approach adopted in naturalistic observation, participant observation is when an evaluator actively gets involved (or participates in) the context they are trying to study. This method involves learning through exposure to or participation in the initiative or programme being evaluated. For instance, an evaluator could participate in a mentorship programme for postdocs to evaluate whether this programme informs participants about their career development opportunities and improves their confidence to apply for lectureships. With this approach, it is important for an evaluator to make other participants aware of the research they are conducting and obtain their consent to participate. As with other qualitative methods, it is important that evaluators consider how their own experiences might influence the data collected and how it is interpreted (for more information on this, see ECU's research and data briefing on **reflexivity**).

Document analysis. The range of documents available depends on the context or population at the centre of the case study, but can include records such as mission statements, annual reports, policy manuals or handbooks, press releases, newspaper articles or blog posts. Document analysis can provide invaluable information about the outcome of an intervention as well as guide the development of other data collection methods (eg what topics to include in a semi-structured interview).

Choosing how to collect data for a case study will depend on that case study's evaluative purpose. There are three main types of case studies (Yin, 1994):

- exploratory case studies aim to answer the 'what' or 'who' questions; for example, if the question is what students do to perform better on their exams, an exploratory case study might ask 'Does a student use any strategies when they read a text?' and 'if so, how often?'
- descriptive case studies aim to describe natural phenomena as they occur, such as which strategies the student used and how the student used them; descriptive case studies are frequently presented in narrative form (McDonough and McDonough, 1997)
- explanatory case studies aim to answer the 'how' or 'why' questions without trying to control the specific group or individual being examined

Case studies can be used to add depth and real-life examples to information about a programme or policy, or apply this descriptive information to generate hypotheses for future evaluations. Where the findings from multiple case studies are brought together, they can be thoroughly compared and evaluated in a cumulative manner.

In general, there are two major strengths of the case study approach (regardless of which type is employed). The first strength is that the researcher conducting the case study is typically responsible for both data collection and its analysis, allowing the two processes to feed into one another fluidly. This allows new lines of inquiry, or extraneous sources of influence, to be incorporated into the evaluation as it is taking place.

One of the key contributors to conducting a plausible case study evaluation is to start data analysis and data collection concurrently. Analysis begins with the first document review, the first interview, or the first observation. (Balbach, 1999)

For example, if an interview with a participant suggests that being on a part-time contract limits the amount of information on promotion processes that they obtain, the researcher can incorporate this as a discussion topic for future interviews, or collect information on participants' contracts through a questionnaire. However, with this strength comes a caveat, as the experiences and beliefs of those conducting the research can frame their interpretation of the data and result in biased data collection. But no research, whether it be qualitative or quantitative, is completely free of bias. The issue of situating the researcher within the research, and how a researcher's own experiences and beliefs can be reframed as an asset, was discussed in a previous ECU research and data briefing on **reflexivity**.

The second major strength of case studies is they are particularly useful for evaluating impact when an initiative is unique. This is because the overarching aim of a case study is to understand a selected initiative or individual as a distinct whole, operating within its particular context. This flexibility also makes case studies a useful approach when an existing initiative is being trialled in a new setting, an unpredictable setting or with a new outcome.

However, focusing on a specific participant or context means that the results of an individual case study cannot be generalised to other contexts or populations. While a single case study may reveal that an initiative has had the desired influence on an outcome, a series of case studies across a variety of contexts and populations is required to determine whether this impact is consistent. If you are aiming to compare your results, the main obstacle in conducting a case study then becomes making sure the information collected is comparable to other case studies evaluating similar initiatives.

Keeping a clear, detailed record of the ways data were collected, how this information was treated during analysis and how these were interpreted in turn are all ways to support similar case studies being conducted in another context or with another population, to add to the body of evidence. As more information accumulates, the stronger the evidence becomes.

The design process and additional resources

There are a number of useful guides on how to design an effective impact evaluation plan (see figure 11), all of which underscore the importance of early planning and making sure the plan sets clear and achievable objectives. Early planning makes it easier to collect data in a systematic and practical manner. It embeds the evaluation of an initiative in the implementation of the initiative itself by clarifying the direction its objectives and facilitating the collection of baseline (pre-intervention) data. In other words, the evaluation of an initiative starts with deciding which methods and measures are going to be used **before** implementing the initiative itself.

A detailed evaluation plan should include:

- background information on the initiative being evaluated (eg what does the initiative aim to achieve? Who is in charge of its implementation? Who will be in charge of collecting and protecting data? Who is the target audience?)

- what questions and indicators will be used in the evaluation (eg survey items, proportions of staff or students)

= how the information obtained will be used and shared with others (eg are the results going to be used to support implementing the initiative full-time? Will the results be used to improve the initiative? How will results be presented to stakeholders and participants?)

The first step is to think about what the results of the evaluation will be used for and set the initiative's goals and objectives.

= **goals:** broad, not time-limited or concrete (eg improving gender balance of senior members of staff; reducing the BME attainment gap among students)

= **objectives:** clear, specific and can be achieved within the project timeframe (eg increasing staff awareness of promotion processes; increasing BME students' feelings of belonging)

An evaluation's objectives are typically centred on the target outcomes of the programme or initiative being implemented. As required in Athena SWAN action plans, evaluation objectives should be SMART (specific, measurable, achievable, relevant and time-bound).

The second step of planning an impact evaluation is developing the questions and indicators that will be used. While the questions used in an evaluation will always need to be tailored to the specific initiative and its context, their development does not always need to be done from scratch (see ECU's guidelines on **writing effective equality surveys** for example). Other available resources include: regional networks, equality and diversity practitioners in other institutions or departments, academic researchers working in social sciences, and so on.

Finally, the third step is to select an evaluation approach. This will depend on the type of questions and indicators being used, as well as what kind of initiative is being evaluated and its target audience. While the advantages and disadvantages of the traditional and alternative methods described in this briefing are a guide for selecting an evaluation approach, it is important to balance available resources (ie time, money, required expertise) with the level of rigour the evaluation requires. More rigorous evaluations such as randomised control trials produce more confident results, but are time-consuming and costly. Thus, it is important to consider how the results will be used, and by whom, when selecting an evaluation method (or methods).

Figure 11: Useful resources for practitioners who want to know more about impact evaluation.

BetterEvaluation.org
This online repository of information is a great starting point for practitioners who are new to evaluation and finding it difficult to select an evaluation method.

Project ECHO® Evaluation 101: A practical guide for evaluating your program
A guidance document describing the individual stages of how to design and implement an impact evaluation plan alongside a programme or initiative. While this guidance document is intended for research on medical programmes, it is easy to follow and can be adapted to equality and diversity initiatives.

OFFA Impact evaluation guidelines
This guidance document on how HEIs can select and conduct impact evaluations of their outreach and widening participation objectives can be adapted and used for evaluations of other equality and diversity initiatives.

NCVO Knowhow Non-profit

The National Council for Voluntary Organisations (NCVO) charity offers a series of webpages that help organisations plan evaluation, develop a theory of change, and collect outcomes and impact data. Resources to help you improve the way you measure impact are also available from **Inspiring Impact** – a UK-wide, ten-year collaboration (between NCVO Charities Evaluation Services and others), which aims to improve impact measurement in the non-profit sector.

Conclusion

The overarching aim of this briefing was to provide equality and diversity practitioners with a starting point for learning about what impact is and the different ways that it can be evaluated. Specifically, this briefing aimed to improve practitioners understanding of impact evaluation methods by presenting them through an equality and diversity lens. As such, the descriptions of the traditional and alternative evaluation methods were described at an introductory level, with more detailed resources presented in figures 9 and 11. Next, this briefing walked practitioners through the design process:

- **Step 1:** state the objectives of the evaluation. These should be clear, specific and time bound.
- **Step 2:** decide what the target outcome of the objectives is and how it will be measured.
- **Step 3:** select which evaluation method is best-suited to showcase change in the target outcome measure. For example, if the outcome is measured by a staff or student survey but participants cannot be tracked across time, a between-groups comparison is more likely to be the appropriate method. In contrast, if the initiative being evaluated only targets a small number of participants, case studies are more likely to shed light on whether there was an impact on the desired outcome.

The inclusion of the main advantages and disadvantages of each method, alongside an outline of the steps involved in the design process and examples situated in an equality and diversity context, provide a comprehensive starting point for practitioners planning an impact evaluation.

References

Balbach, E; the California Department of Health Services (1999). Using case studies to do program evaluation. **www.case.edu/affil/healthpromotion/ProgramEvaluation.pdf**

Crawford, C; Dytham, S; Naylor, R (2017). *The evaluation of the impact of outreach*. Office for Fair Access, UK. **www.offa.org.uk/wp-content/uploads/2017/06/Standards-of-Evaluation-Practice-and-Associated-Guidance-FINAL.pdf**

Earl, S; Carden, F; Smutylo, T (2001). *Outcome mapping: building learning and reflection into development programs*. International Development Research Centre, Canada. **www.idrc.ca/en/book/outcome-mapping-building-learning-and-reflection-development-programs**

ECU (2011). *Effective equality surveys: exploring the staff and student experiences in higher education institutions*. London, UK. **www.ecu.ac.uk/publications/effective-equality-surveys**

ECU (2015). *ECU's Athena SWAN Charter: guide to processes, May 2015*. London, UK. **www.ecu.ac.uk/wp-content/uploads/2015/05/Athena-SWAN-Charter-Post-May-2015-guide-to-processes.pdf**

ECU (2016). *Research and data briefing 1: working with data*. London, UK. **www.ecu.ac.uk/guidance-resources/using-data-and-evidence/working-with-data**

ECU (2017). *Reflexivity: positioning yourself in equality and diversity research*. London, UK. **www.ecu.ac.uk/publications/reflexivity-positioning-yourself-in-equality-and-diversity-research**

Field, A (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage Publications.

Field, A; Miles, J; Field, Z (2012). *Discovering statistics using R*. London, UK: Sage Publications.

Gabarino, S; Holland, J (2009). *Quantitative and qualitative methods in impact evaluation and measuring results*. Governance and social resource development centre, UK. **www.gsdrc.org/docs/open/eirs4.pdf**

HEFCE; SFC; HEFCW; DfE (2017). *REF 2021: A guide for research users*. London, UK.

McDonough, J; McDonough, S (1997). *Research Methods for English Language Teachers*. London, UK: Arnold.

The New York Academy of Medicine; GE Foundation; The NYS Health Foundation (2017). Project ECHO® Evaluation 101: A practical guide for evaluating your program. New York, NY. **https://nyshealthfoundation.org/wp-content/uploads/2017/12/project-echo-evaluation-guide.pdf**

United States General Accounting Office, Program evaluation and methodology division (1990). *Case study evaluations*. Denver, CO. **www.gao.gov/assets/80/76069.pdf**

Yin, R (1994). *Case study research: Design and methods* (2nd ed.). Beverly Hills, CA: Sage Publishing.

Equality Challenge Unit (ECU) supports higher education institutions across the UK and in colleges in Scotland to advance equality and diversity for staff and students.

ECU provides research, information and guidance, training, events and Equality Charters that drive forward change and transform organisational culture in teaching, learning, research and knowledge exchange. We have over fifteen years' experience of supporting institutions to remove barriers to progression and success for all staff and students.

We are a registered charity funded by the Scottish Funding Council, the Higher Education Funding Council for Wales and through direct subscription from higher education institutions in England and Northern Ireland.

@EqualityinHE
www.ecu.ac.uk