

Types and strength of evidence: assessing sources, and producing your own evidence

Rain Sherlock
Dr Eliza Kozman
May 2023

Overview of session

1

Generating your own evidence

- What are the methods?
- TASO examples
- Strengths/limitations

2

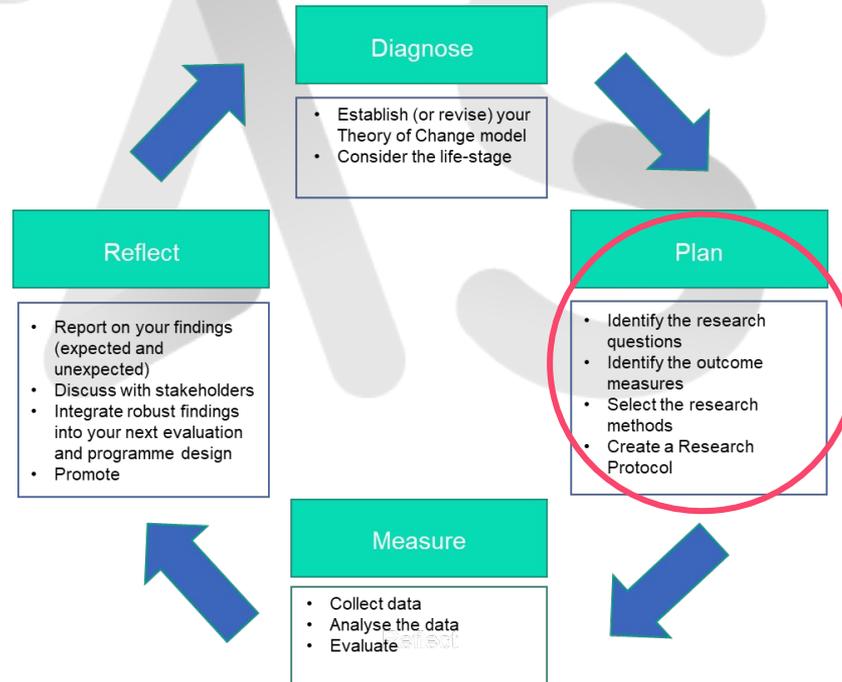
Assessing evidence

NO MATHS!

Generating your own evidence



The evaluation cycle



What is your research question?

Primary Research Question

Causal impact of your evaluation

Did [scheme] increase [main outcome] among [group]?

Did Summer School attendance improve enrolment rates among participants?

Secondary Research Question

Focus on specific **groups** or **intermediate outcomes**

Did [scheme] increase [main outcome/ secondary outcome] among [group/subgroup]?

Did Summer School attendance improve enrolment rates among estranged students?

ACTIVITY: Research question (4m)

- Consider a specific intervention you run at your institution
- Discuss with your neighbour your primary research questions
- Consider:
 - Who will use the findings and how?
 - What do stakeholders need to learn from the evaluation?
 - What questions will you be able to answer and when?

Outcome measures

“I’ll know [**outcome reached**] when I see [**indicator**]”



Observable indicators are those we can build into the evaluation and control out; e.g. demography, observed behaviour, measured attitudes.



Unobservable indicators are those we can’t observe and therefore can’t build into the evaluative model; e.g. motivation, unobserved behaviour, unmeasured attitudes; anything that influences the outcome that we don’t know about or can’t measure.

Outcome measures

1. Core impact (e.g. A level attainment, university acceptances, continuation)
2. Interim or proxy outcome (e.g. GCSE selections, sign-ups to events)
3. Validated scales (e.g. from academic research, externally-administered tests)
4. Self-report objective (e.g. actual knowledge)
5. Self-report subjective (e.g. perceived knowledge)

TASO Common Outcome Measures

TASO: Common Outcome Measures

Year group	Objectives	Indicators for Process Evaluation	Indicators for Impact Evaluation (short-medium term)	Indicators for Impact Evaluation (long-term)
<i>Which age group or Key Stage are you engaging with?</i>	<p><i>What is your intervention trying to achieve?</i></p> <p><i>You will have identified your objectives through your Theory of Change</i></p>	<i>Which indicators will enable you to test whether your interventions are being delivered/implemented as intended and whether this process can be improved?</i>	<i>Which indicators will enable you to identify whether you have achieved your intermediate outcomes?</i>	<i>Which indicators will enable you to identify whether you have met/achieved your long-term goals?</i>
KS3 (Years 7-9)	<ul style="list-style-type: none"> Improving metacognitive skills to raise attainment IAG on how to progress to HE Increasing awareness of subject options at HE Raising aspirations Engaging parents 	<ul style="list-style-type: none"> Was the programme delivered as intended? Were students targeted correctly? Did students attend? Which students attended? Was the content delivered as intended by deliverers? (academics, ambassadors etc.) Participant experience as measured by evaluation survey Experience & perceptions of stakeholders e.g. key influencers such as teachers, ambassadors, project leads, academics 	<ul style="list-style-type: none"> Improvement in attainment as measured by in-school exam scores e.g. SAT scores Improved metacognitive skills as measured by the Junior Metacognitive Awareness Inventory (JMAI) Students' self-reported understanding of higher education 	<ul style="list-style-type: none"> GCSE attainment (through NPD) A level attainment (through NPD/HEAT) Progression to university (HEAT) Progression to research intensive/ highly selective HEIs (HEAT)

WP questionnaire

The WP Questionnaire

Sense of belonging (prospective)

Sense of belonging

Academic self-efficacy (prospective)

Metacognitive strategies

Study strategies

Critical engagement with information

University expectations and knowledge

Rapid review:

Intermediate outcomes for higher education access and success

Authors: Hannah Thomson, Lauren Bellaera, Sonia Ilie, Konstantina Maragkou

November 2022

GUIDANCE FOR USING THE WIDENING PARTICIPATION QUESTIONNAIRE SCALES

This document will help you understand how to use the validated questionnaire scales when trying to evaluate a student access or success programme. It is important that the scales are implemented as suggested here, to help with consistency around the data.

BEFORE YOU START

What is a validated questionnaire scale?

A validated questionnaire scale is a set of questionnaire items (statements or questions), which together capture a specific outcome of interest. A validated scale has undertaken a multi-step process to ensure it captures the specific outcome it sets out to capture, does so reliably, and generates good quality data.

Validation is a complex process that requires a large number of steps. Read our [method note](#) to understand how we've gone about validating the scales in this widening participation questionnaire.

Why would I use a validated scale?

When trying to evaluate a student access or success programme, using a low-quality questionnaire scale can be a barrier to gaining meaningful insights from your work. While many questionnaire scales are available, information about their quality is not always forthcoming or complete, and a question mark remains over their applicability to the samples of learners you might be working with. Using a validated questionnaire scale in your evaluation will contribute to the quality of your data, and therefore to the quality of the insights you can draw about your programme from your evaluation.

When would I use one of these validated scales?

Use one of these validated scales when you are trying to evaluate a student access or success programme and the expected outcomes of that programme align with the outcomes which our scales cover. Instead of designing a new questionnaire, buying one, or using something about which you do not have a lot of information, choose the validated scale that aligns with your outcome(s) of interest, then include this measure in whichever survey data collection approach you take as part of your evaluation. Finally, analyse data and obtain your evaluation result feeling reassured that you have used a good-quality measure which has not negatively affected the ability of your evaluation to offer insights about your programme.

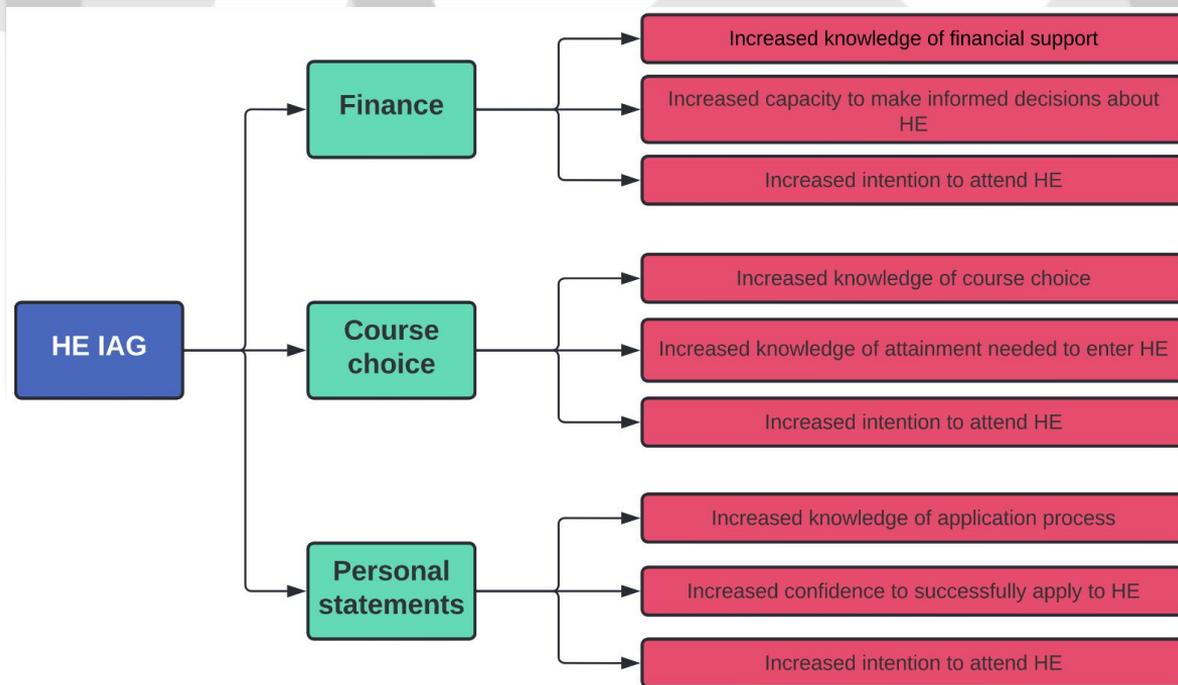
Are these the only scales I need?

Your student access or success programme may seek to affect a large number of outcomes. The validated scales provided here only capture some of those outcomes. We call these **intermediate outcomes**, because they are often thought of as precursors of later outcomes such as access to higher education, continuation, progression, or degree outcomes. There may be other outcomes that will be relevant to your programme, and as such, other questionnaire scales that you should use. Future guidance will focus on how you may go about validating scales yourself, for your own use. In the meantime, you can find guidance on survey design and validation on the [TASO website](#).

Do I need to use all the scales, at the same time?

Not at all. You should only use the scale, or scales, that are relevant to the outcomes of the programme you are looking to evaluate. The scales are here to strengthen the quality of your evaluation, but it is your own programme that will be driving the focus of your student access and success work and the outcomes that you are most interested in.

Mapping Outcomes & Activities Tool (MOAT)



ACTIVITY: Outcomes (4m)

- Consider an appropriate outcome indicator for the intervention you want to evaluate.
- Could it be improved? How?

Impact and process evaluation

- **Impact Evaluation (this session)**
 - Focused on the impact of the intervention
 - Does it work?
 - Helps decide whether an intervention should be adopted, continued, or modified for improvement.
- **Implementation and Process Evaluation (later)**
 - Focused on whether the intervention is being implemented as intended and whether the underlying assumptions hold

Standards of evidence

Office for Students

Type 1: Narrative	The evaluation provides a narrative and a coherent theory of change to motivate its selection of outreach activities in the context of a coherent outreach strategy
Type 2: Empirical Research	The evaluation collects data on outcomes and impact and reports evidence that those receiving an intervention have better results, though this does not establish any direct causal effect
Type 3: Causality	The evaluation methodology provides evidence of a causal effect of an intervention



We've got quite a lot of this



Some of this, and getting better

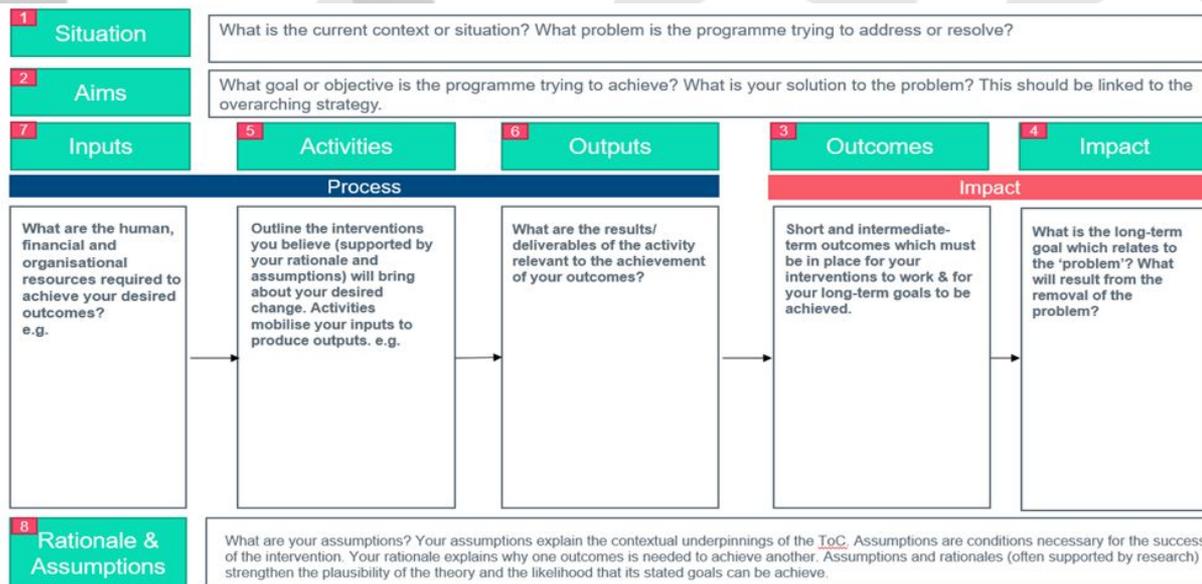


Missing piece of the puzzle

OfS standards of evidence - Type 1

Type of evidence	Description	Evidence	Claims you can make
Type 1 - narrative	The impact evaluation provides a narrative or a coherent theory of change to motivate its selection of activities in the context of a coherent strategy.	Evidence of impact elsewhere and/or in the research literature on access and participation activity effectiveness or from your existing evaluation results.	We have a coherent explanation of what we do and why our claims are research-based.

Theory of Change



Type 1 considerations

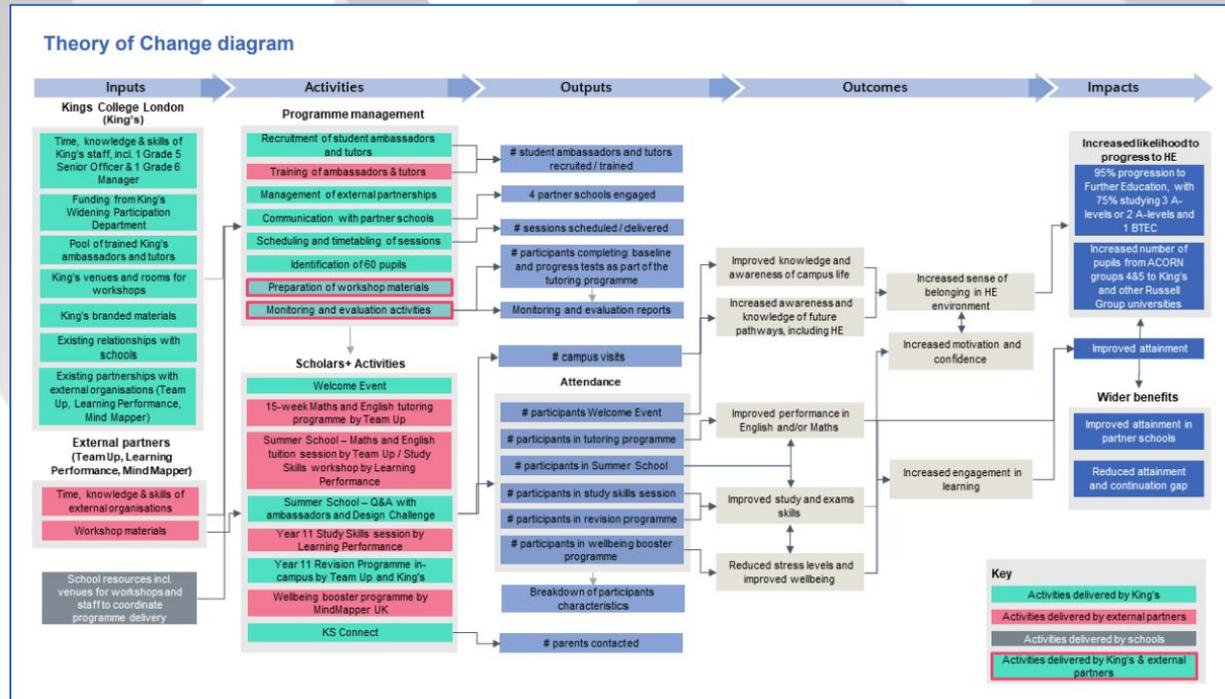
Strengths

- A coherent explanation of why an intervention should work
- Clear base in existing evidence

Limitations

- No actual data on impact of your intervention

TASO example: attainment-raising ToCs



OfS standards of evidence - Type 2

Type of evidence	Description	Evidence	Claims you can make
Type 2 – empirical enquiry	The impact evaluation collects data on impact and reports evidence that those receiving an intervention have better outcomes, though does not establish any direct causal effect.	Quantitative and/or qualitative evidence of a pre/post intervention change or a difference compared to what might otherwise have happened.	We can demonstrate that our interventions are associated with beneficial results.

Type 2: association with better outcomes

- Move towards making more robust claims
- Taking part is associated with better outcomes

Pre-post testing

Rate on the following Likert scale:

How happy do you currently feel?

Very unhappy	Unhappy	Slightly unhappy	Neutral	Slightly happy	Happy	Very happy
1	2	3	4	5	6	7

Very important intervention



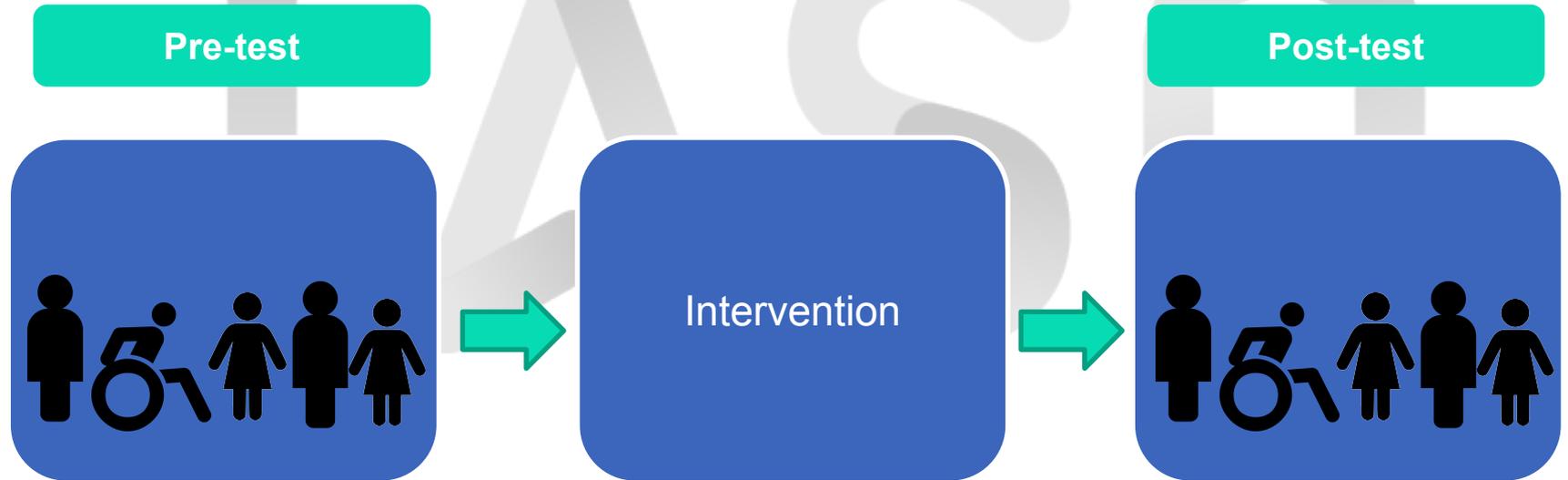
Pre-post testing

Rate on the following Likert scale:

How happy do you currently feel?

Very unhappy	Unhappy	Slightly unhappy	Neutral	Slightly happy	Happy	Very happy
1	2	3	4	5	6	7

Pre-post testing



TASO example: summer schools pre-post

- Assessing the impact of university summer schools.
- As part of this evaluation, included a pre-post element, alongside a more robust randomised evaluation



TASO example: summer schools pre-post

EXAMPLE

Table 8: Estimated effects for the outcomes of interest

Outcome	Pre-summer-school mean	Post-summer-school mean	t statistic	Estimated effect (Cohen's <i>d</i>)
Likelihood of progressing to HE (7-point Likert scale)	6.42	6.65	2.16*	0.20
Self-efficacy relating to HE application (5-point Likert scale)	3.68	3.87	2.72**	0.24
Self-efficacy relating to post-entry success (5-point Likert scale)	3.94	4.01	1.18	0.10
Compatibility of HE with social identity (5-point Likert scale)	3.73	3.94	2.76**	0.23
Perception of financial barriers to HE (5-point Likert scale)	2.99	3.55	6.49***	0.58
Perception of knowledge barriers to applying to HE (5-point Likert scale)	3.13	3.94	9.06***	0.81

Notes: n = 142

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Pre-post considerations

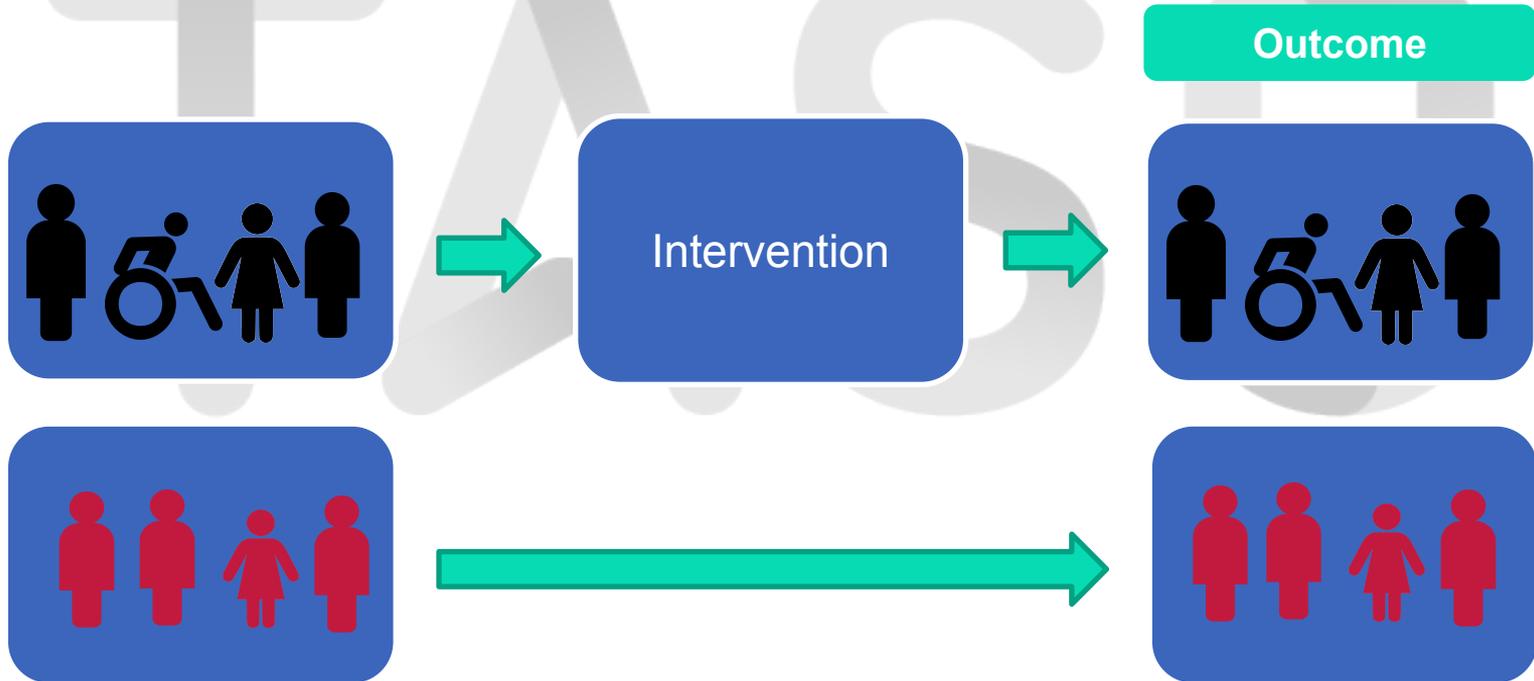
Key strengths

- Can be straightforward to implement, particularly with survey data
- Easy to interpret - the change over time

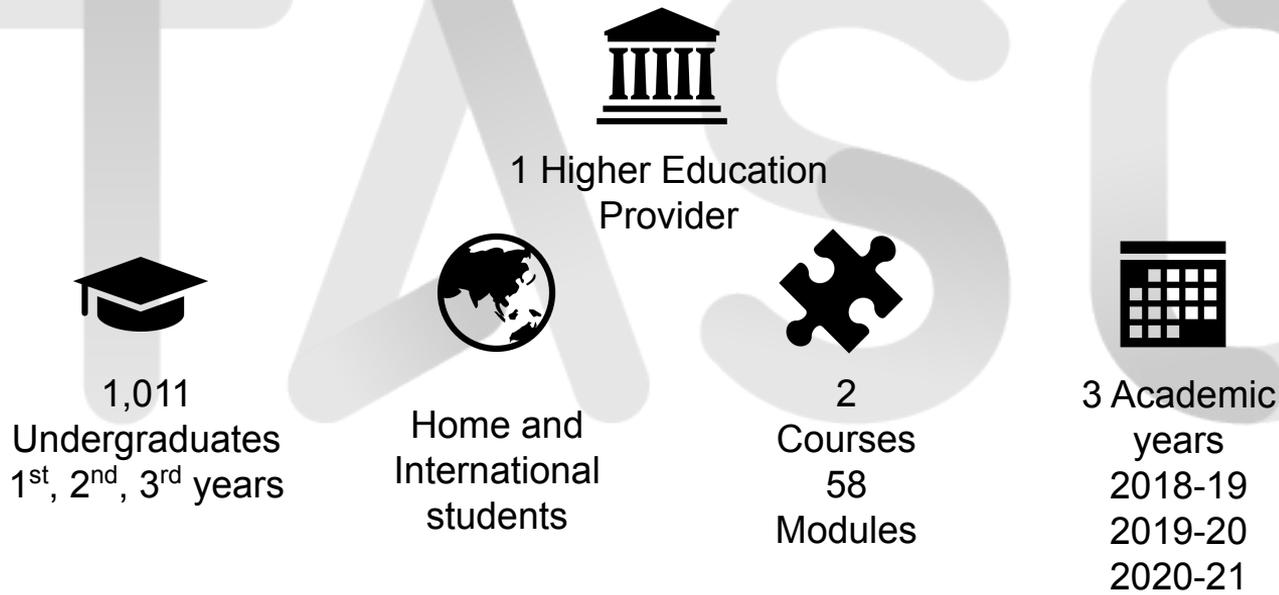
Key limitations

- Assumes that if the intervention didn't happen the outcome would not change - doesn't consider other possible drivers of change
- No comparison to other groups
- The only outcomes that can be measured are those that can be collected at multiple time-points
- Testing effects can be an issue

A basic comparison

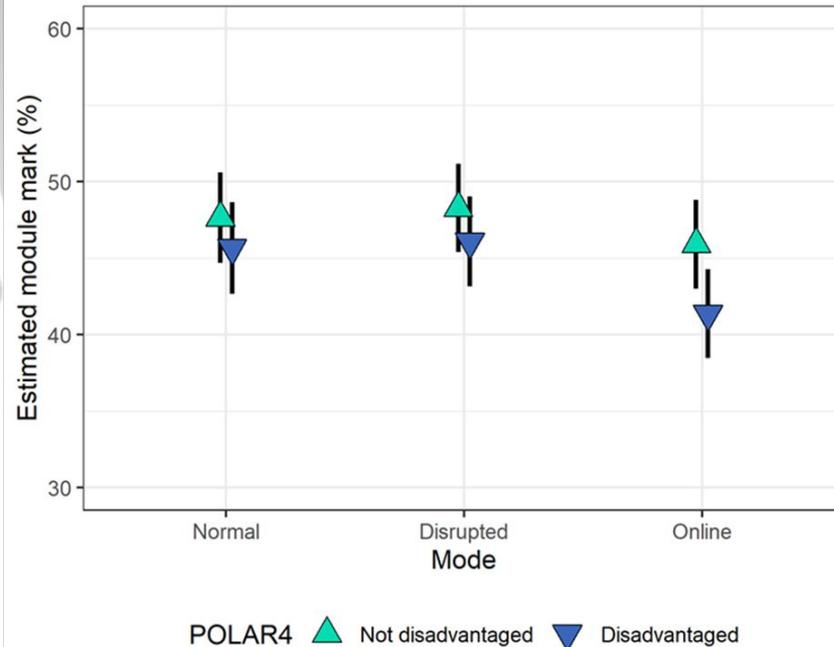


TASO example: teaching and learning analysis



TASO example: teaching and learning analysis

Mode	Gap
Normal	2.1 pp
Disrupted	2.5 pp
Online	4.8 pp



EXAMPLE

Basic comparison considerations

Key strengths

- Can use existing institutional data and historical data sets
- Can show correlation between intervention and outcomes

Key limitations

- Basic comparison to other people who didn't take part
- Could be comparing apples with oranges
- Not able to make strong claims about impact

OfS standards of evidence

Type of evidence	Description	Evidence	Claims you can make
Type 3 – causality	The impact evaluation methodology provides evidence of a causal effect of an intervention.	Quantitative and/or qualitative evidence of a pre/post treatment change on participants relative to an <u>appropriate control or comparison</u> group who did not take part in the intervention.	We believe our intervention causes improvement and can demonstrate the difference using a control or comparison group.

Type 3: why we need it

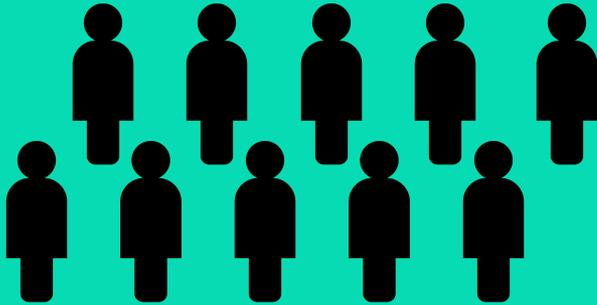
- **Correlation doesn't mean causation**
 - Type 2 evidence is useful in building our understanding.
 - Type 3 methods can uncover the fuller picture.



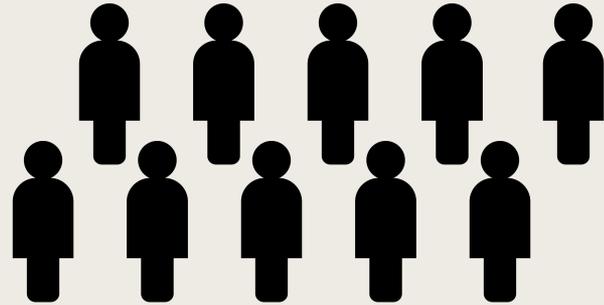
© marketoonist.com

Example: WP activity

Apply for WP activity

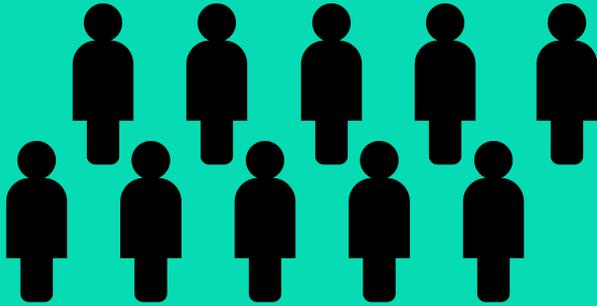


Do not apply



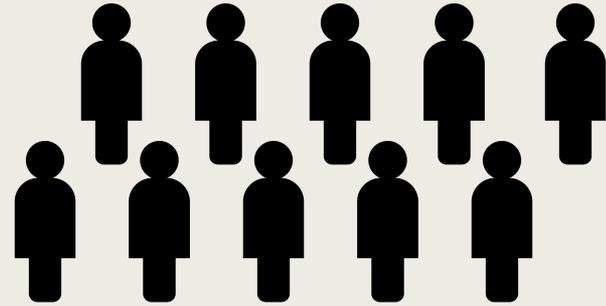
Example: WP activity

Apply for WP activity



80% enrol in HE

Do not apply



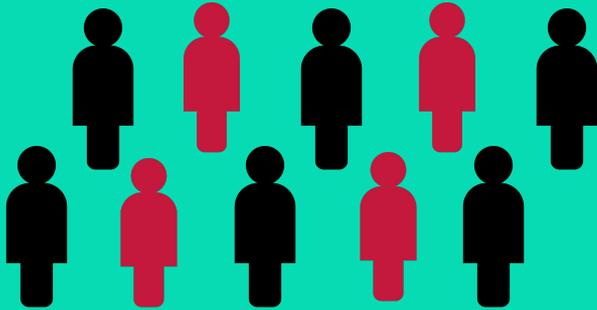
50% enrol in HE

Differences are there between groups?

- **Demographic differences?** (e.g. gender, prior attainment, location)
- **Other differences?** (e.g. family support, individual motivation, other barriers)?

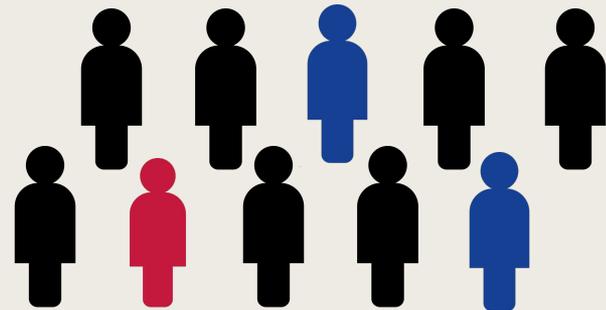
Selection bias

Apply for WP activity



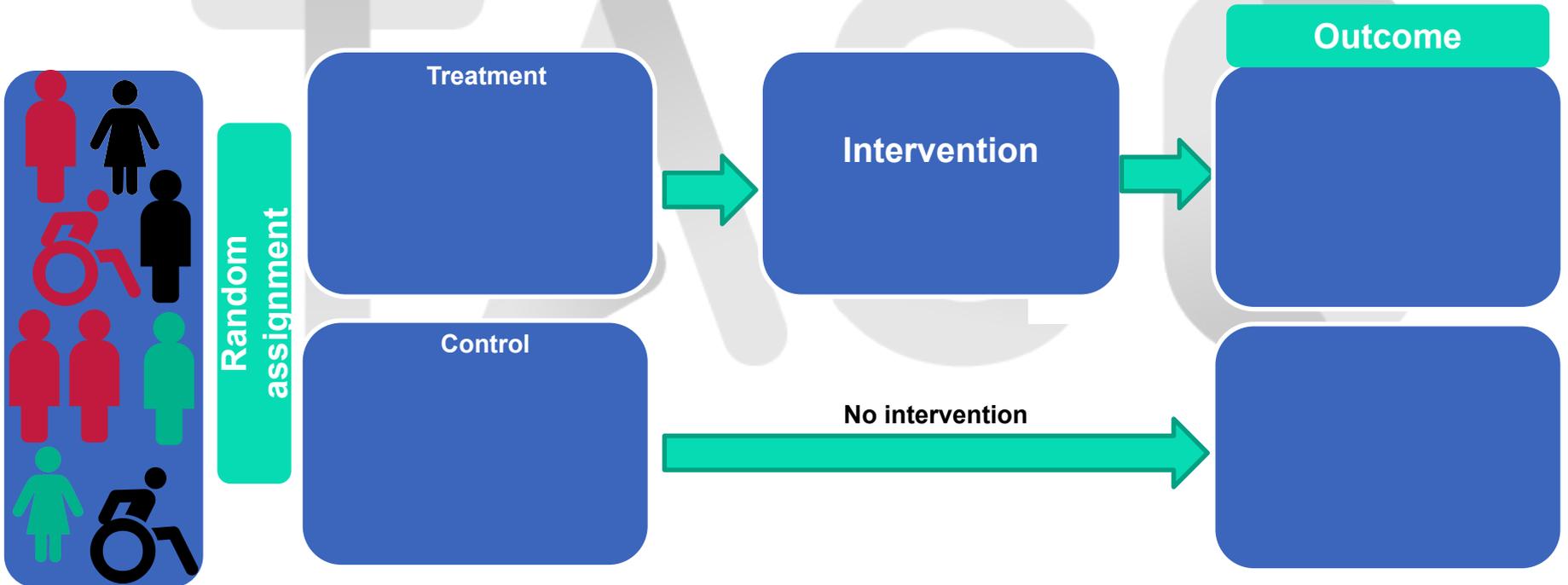
80% enrol in HE

Do not apply

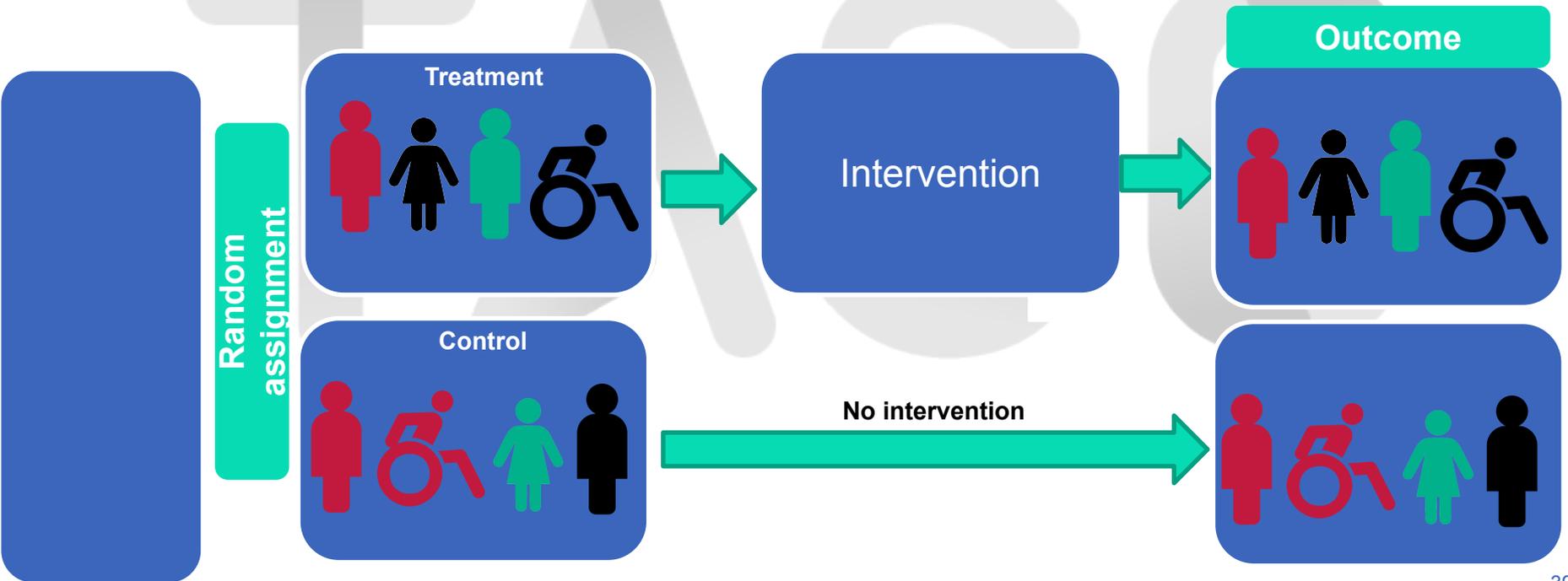


50% enrol in HE

Random comparison group



Random comparison group



RCT Considerations

Strengths

- A Type 3 method under the OfS Standards of Evidence
- Ideal tool for causal inference
- Helpful in determining whether an intervention / public policy works
- Typically have high internal validity (to be defined later!)

Limitations

- Can be time consuming and expensive
- Hawthorne effect and other internal validity concerns
- Sometimes difficult to randomize participants
- Can require large sample
- External validity considerations

TASO example: summer schools RCT

- University partners
 - Pre-16
 - Post-16
- Summer schools 2021 - online delivery
- Summer schools 2022 - face-to-face delivery



TASO example: online summer schools RCT

- Summer schools are oversubscribed.
- Eligible applicants are **randomly** allocated to either receive a place or not receive a place.



TASO example: online summer schools RCT

Outcome	Intervention mean	Control mean
Likelihood of going to HE (7-point likert scale) (N = 342)	6.60	6.60
Applied to HE (binary yes/no) (N = 295)	0.94	0.91

EXAMPLE

Quasi-experimental designs

- No randomisation
- But try to simulate something like an RCT using data we already have
- For examples see:
 - Race equality gaps report
 - Multi-intervention outreach and mentoring report

Finding comparison group data

Source	Good for...
UCAS Outreach Evaluator	<ul style="list-style-type: none"> - Comparing potential applicants with national average, a control group and “your competition” - Application to acceptance
Jisc	<ul style="list-style-type: none"> - Comparing students to a matched comparison group within the HESA dataset
Institutional data	<ul style="list-style-type: none"> - Comparing students on different courses or who do and don't engage with a service/programme
Survey (bespoke or part of wider study)	<ul style="list-style-type: none"> - When you are looking at softer outcomes

ACTIVITY: Research methods (10m)

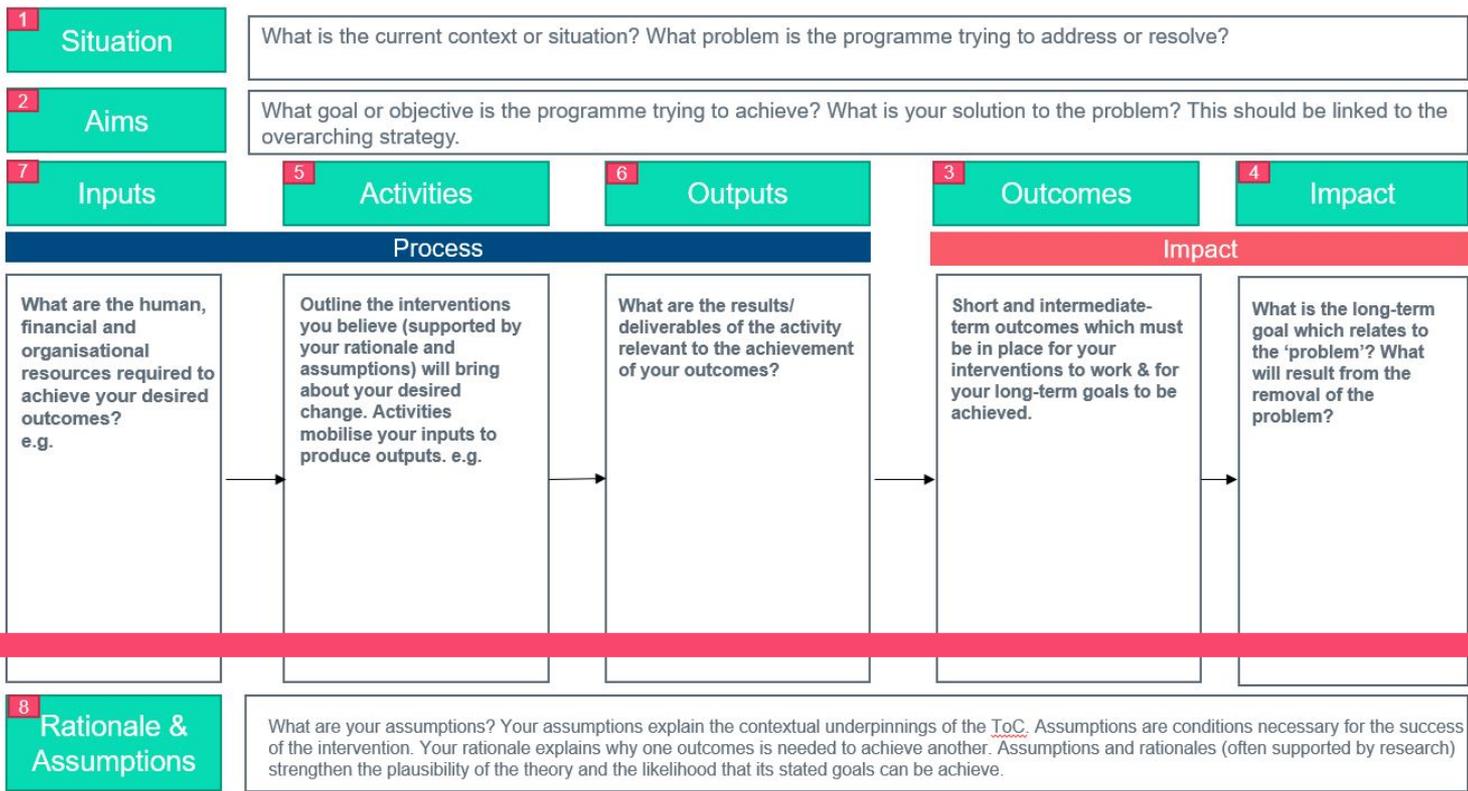
- Split into groups of 3-4 to discuss
 - Pre-post
 - Comparison groups
 - Randomisation
- Discuss possible research designs for your evaluation
- What are three key steps you can take to progressing this in your institution?
- **TASO team members will circulate**

Questions?

TASO

Assessing evidence





Rationale and assumptions

- How effective do we think it will be?
- Why?
- For who?
- In what context?
- ...

Rationale and assumptions

STAND ON THE SHOULDERS OF GIANTS!



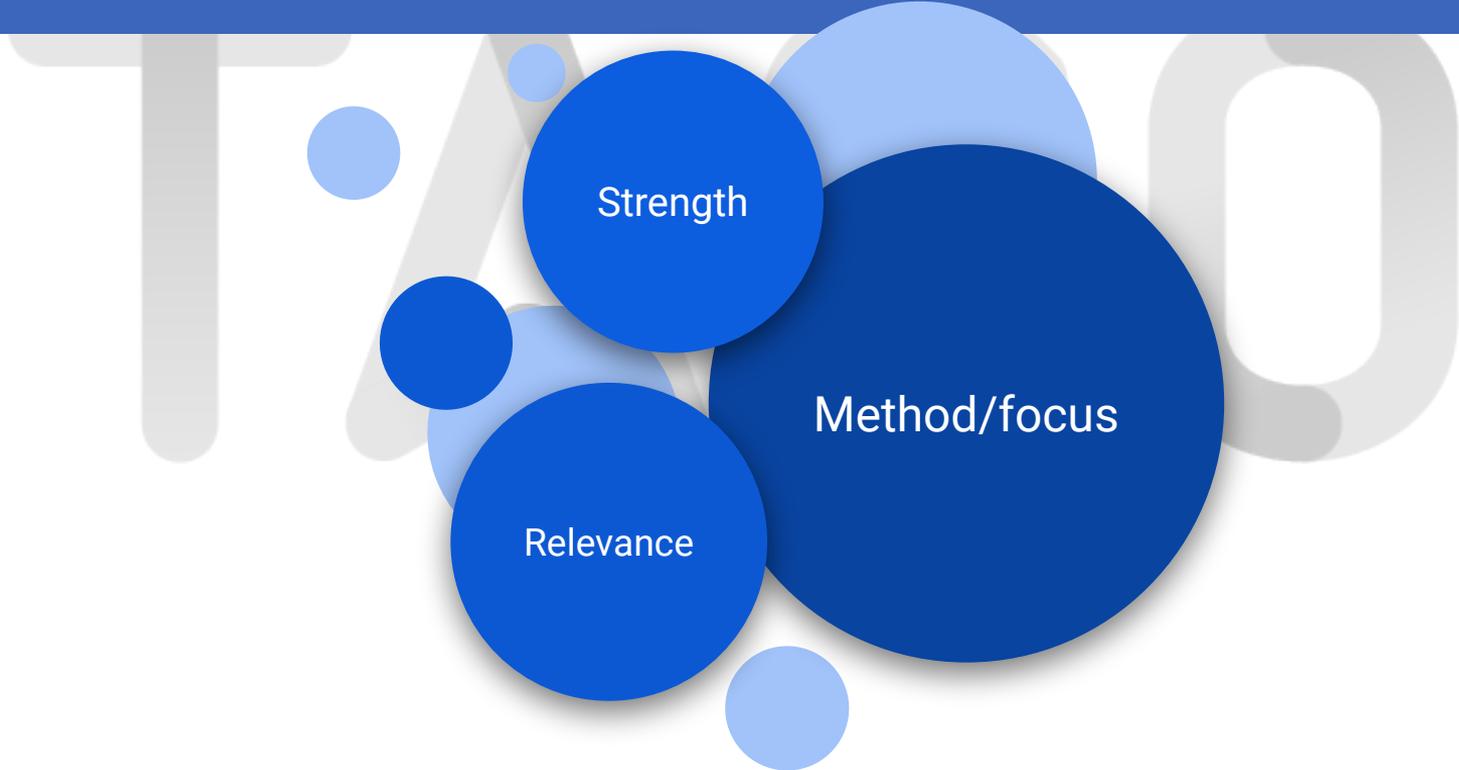
Where to get evidence?

- TASO reports and toolkit
- OfS reports
- Effective practice
- EORR - and our review
- Uni Connect
- Academic journals
- NBER
- ERIC
- EEF
- ...

The screenshot shows the TASO website interface. At the top left is the TASO logo and tagline. To the right is a 'Sign up for our newsletter' button and social media icons for Twitter and LinkedIn. Below the header is a navigation menu with 'RESEARCH', 'EVIDENCE & EVALUATION', 'GET INVOLVED', 'NEWS & BLOG', and 'ABOUT'. A search icon is on the far right. A breadcrumb trail reads: Home > Evidence & Evaluation > Evidence toolkit > Financial support (post-entry). The main heading is 'Financial support (post-entry)'. Below it is a paragraph: 'Financial support includes grants, bursaries, scholarships and fee-waivers. When offered after entry to higher education, it is designed to help students succeed on-course by alleviating the financial costs of studying.' There are four filter tabs: 'Completion', 'Progression to own institution', 'Wellbeing', and 'Post-entry to HE'. Below this is a grid of four evidence metrics:

Metric	Value
Cost	High cost (represented by three 'f' icons)
Impact on aspirations / attitudes	Small positive impact (represented by a plus icon)
Impact on behaviour / outcomes	Small positive impact (represented by a plus icon)
Strength of evidence	Emerging evidence (represented by three stars)

All evidence is not equal



No such thing as ‘good’ evidence

- **Relevance**
 - How does the evidence apply to the problem we are to address?
- **Method/focus**
 - What is the research question?
 - Is the research/evaluation trying to focused on impact or on something else?
- **Strength**
 - Given the research question and focus, is the design strong?

But some is more useful than other

- **For the purpose of APPs, what we want to understand is impact i.e. how well interventions are working**
- Some sorts of evidence are better for this purpose
- Consider the limitations outlined earlier

Internal/external validity

Internal

- To do with the methods and design - how confident are we in the findings? Are the conclusions correct?
- Common threats include:
 - **Selection bias** - systematic differences between groups
 - **Testing effects** - e.g. participants changing their behaviour because they are observed
 - **Attrition** - loss of data e.g. low survey responses

External

- How generalisable is the evidence? Is it useful to us?
- Common threats include:
 - Focus on particular samples
 - Artificial settings or interventions which don't translate
 - Cultural or geographical differences
 - Contextual factors e.g. pandemics

Top tips for assessing evidence

Relevance

- What is the intervention?
- What is the focus group?
- What is the setting? country/institution type

Method/focus

- In particular, is there a comparison between people who did and didn't get 'the thing' we are interested in?

Strength

- What is the sample size?
- What is the outcome measure?
- If there is, are there other possible reasons why we might see a difference in outcomes?
- What are the stated limitations (read the paper, but be prepared to point out ones they've missed!)

Most common limitations

- No comparator or problems with the comparison group
- Outcomes are self-reported or not very strong in other ways (e.g. short-term)
- Small samples
- Not really about impact!

Activity (10 minutes)

- Consider a specific intervention you run at your institution
- Consider
 - What is the evidence you use to underpin it?
 - What are the weakness of this evidence?
 - What evidence would you ideally want...can you generate it?

One swallow does not make a summer

Level	Strength of evidence	What this means
4	Strong evidence	5 or more pieces of OfS Type 3 evidence from the UK
3	Medium evidence	3 or more pieces of OfS Type 3 evidence from the UK
2	Emerging evidence	3 or more OfS Type 2 evidence sources from the UK and/or 3 or more Type 3 evidence sources from outside the UK
1	Weak evidence	Any other number or combination of studies

Research protocols

- A Research Protocol is a written document that describes the overall approach that will be used throughout your initiative, including its evaluation.
- **See our website for open access to our protocols**



Takeaways

- There are a range of ways you can improve your evaluation practice
- Hopefully we have given you some ideas for how you can start generating more evidence and some practical examples
- Please come to us with questions or ideas!

Questions?

TASU

Thank you