What Works *for* Children's Social Care

# Evaluation Pipeline

Version: 1.0
Updated: April 2022

In line with WWCSC's overarching research principles, we expect all our evaluations to adhere to our four principles: they need to be focused on impact, bear in mind the nuance of social care, be useful and help empower the profession.

However, we do not expect all our programmes and evaluations to be at the same stage of development. The aim of this document is to:

- Describe and define the different types of evaluations we support
- Explain how we assign each new programme to a type of evaluation
- Clarify the criteria used when deciding whether a programme and evaluation should be scaled up.

This document is intended for:

- Delivery Partners, i.e. the organisations and people who deliver our programmes.
- Evaluators, i.e. the organisations and people who independently evaluate these programmes.
- WWCSC staff, i.e. the people who manage and promote our work. They include: Programme Managers, Researchers, Practice Development Managers, Policy Managers and Operations Managers. .

We are publishing this document to:

- Clarify our decision-making
- Anticipate challenges which arise through the course of evaluations
- Help all parties make important decisions about the design of our evaluations

This note outlines the fundamental principles that underpin our Pipeline. More details are included in our Evaluation Guidance, which can be found at the bottom of this page.

This document is intended to be used throughout:

- The application stage (call for delivery partners)
- The set-up stage (i.e. after the award stage and before the publication of the protocol)

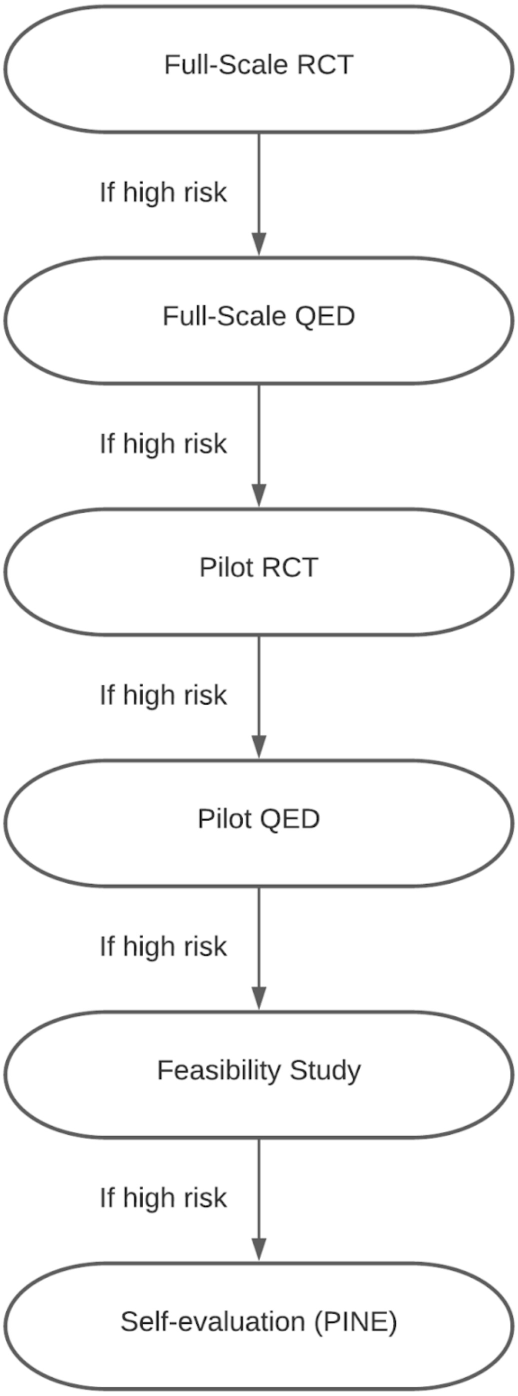
**Feedback and improvements**

Our intention is to update this document on a regular basis to make sure it reflects our latest thinking as well as the feedback we have received. If you have any questions or suggestions, please email: arnaud.vaganay@whatworks-csc.org.uk

| Version | Date | Changes |
|---|---|---|
| V1 | 01.04.2022 | – |
| | | |

## 1. Overview of the Evaluation Pipeline

Our funding process starts with the launch of a call for delivery projects, which typically include priority research topics (although we consider all interventions that aim to improve outcomes in all areas of life for children and young people with a social worker). We then identify the projects most likely to achieve the intended impacts for social workers, children and families, using a decision-making framework. Following that, all parties decide how the project is going to be evaluated. Our Pipeline includes six types of evaluation, which are described and defined in section 2.



Our Pipeline follows a number of principles. We appreciate that some of these principles will at times conflict. When this happens, we are guided by our wider mission and values to find a way forward.

**Replication**

We exist to find out what works, for whom, under what circumstances and why. These questions are best answered through evidence syntheses. This requires a certain rigour in how we scope our evaluations, analyse the evidence and report findings. Because of this mission, we are unlikely to fund evaluations that use untested methods, processes or outcomes, or evaluations that cannot be easily replicated.

**Simplicity**

Our pipeline is linear. Whilst experienced evaluators may find this presentation simplistic, we believe that it is transparent, understandable, and fair. Accordingly, we assess all new projects entering our Evaluation Pipeline following the same sequence and based on the same risk factors (see below), but with enough agility to enable rapid decisions in simple situations. We aim to periodically review these factors, as we learn more about our evaluations.

**Risk**

All trials face two interrelated risks:

- A risk that the intervention will not achieve its objectives (e.g. reuniting participating children with their parents), and
- A risk that the evaluation will not meet its own objectives, (e.g. providing an unbiased estimate of the impact of the programme on said reunifications).

We assess these risks in the early stages of the trial (funding stage, set-up phase) by interrogating the theory of change (ToC) underpinning the intervention, and the ToC underpinning the evaluation. Whilst most trials have an explicit ToC for the intervention being evaluated, few have an explicit ToC for the evaluation. Examples of theories in a given evaluation include:

- social workers refer all eligible beneficiaries, and only those, to the programme being evaluated,
- all participants remain in the group to which they have been assigned (that is, treatment or control) for the duration of the trial,
- data quality is high
- attrition does not exceed a certain level.

These risks affect all parties: if the objectives of the evaluation are unlikely to be met, time, data, resources and goodwill are likely to be wasted. Thus, it is important that all parties assess these risks, using risk factors listed in this document. An evaluation design that carries a high risk should not be undertaken. Our pipeline indicates what design should be considered next.

### Scale

Our preference is to support evaluations where the number of beneficiaries is large enough to detect a statistically significant impact, when there is one. However, we will consider supporting other types of evaluations if the intervention is promising but not ready for a full-scale evaluation.

Projects and evaluations not achieving their full scale will be considered for scaling up using criteria listed in section 4 of this document. Scaling up refers to two different but closely related dimensions, namely an increase in the reach of the programme and a progression from one type of evaluation (e.g., pilot RCT) to the next (e.g., full-scale RCT). These two dimensions go hand-in-hand: we are unlikely to fund a progression from pilot to full-scale evaluation without a significant increase in the programme's reach, and vice-versa.

### Causality

We rank evaluation designs in terms of their ability to generate strong causal claims. Accordingly, our preference is to support Randomised Controlled Trials (RCTs). However, we will consider other evaluation designs if the consensus is that an RCT is not ethical, feasible or cost-effective.

We review the literature to identify the highest level of pre-existing evidence achieved by a given programme. For example, programmes for which the highest level of existing evidence is a pilot RCT will be first considered for a full-scale RCT (using the above-mentioned definitions of 'pilot' and 'full-scale').

### Complexity

We are interested in supporting both simple and complex interventions; however, we recognise that most of the interventions in our portfolio are complex. We define intervention complexity in terms of the number of components, outcomes, stakeholders and interactions.

In line with our mission, our preference is to fund evaluations that can be easily analysed and explained, such as RCTs. A complex evaluation is one that is based on more difficult assumptions and/or requires more data, analysis, or samples.

We recognise that complex interventions often require complex evaluations.

### Summative questions

We expect our evaluations to answer *both* summative and formative questions. Summative questions include:

- Is there a difference in outcome(s) between the different groups/interventions?
- If so, how likely is this difference to be causal?
- To what extent is the theory of change validated?
- Was the evaluation executed as intended?

The emphasis in summative/formative questions is expected to vary according to a number of factors. We expect a greater focus on summative questions when:

- The intervention is simple and/or fully formed;
- The evaluation design is simpler and/or more developed.

### Formative questions

Formative questions include:

- How could the intervention be improved?
- How could the delivery of the programme be improved?
- How could the evaluation be improved?
- E.g., Is version A of the recruitment letter more effective than version B?
- E.g. Is outcome A more reliably collected than outcome B?

The use of rapid-cycle design and testing is strongly encouraged, especially in cases where a full-scale evaluation is considered high-risk.

### Proportionality

The resources allocated to an evaluation, the rigour of its management, the amount of fieldwork and the length of the final report should be proportionate to the investment made in the programme.

### Consensus

Our preference is to make decisions by consensus. Delivery partners, evaluators and WWCSC should be involved in all decisions that affect them. Where possible, advice should be sought from a wide range of sources to minimise groupthink. Potential conflicts of interests and risks of bias should be acknowledged.

### Open-mindedness

We want to work with organisations and people who either share our values, or are open-minded enough to try our approach. We accept and respect other approaches to improve children's outcomes, including where they are not aligned with ours.

## 2. Types of evaluations

We support six types of evaluation.

### Full-scale RCTs

All studies are first considered for a full-scale RCT, unless there is a consensus that this option carries a high risk, using the risk factors presented in the document.

A full-scale RCT is primarily summative:

- The aim of the impact evaluation is to generate a precise impact estimate. It is designed to achieve: sufficient statistical power to detect a realistic minimum detectable effect size, low attrition, and maximum internal validity.
- The aim of the Implementation and Process Evaluation is to evaluate the theory of change.
- It should be clear about the parts of the evaluation that were not executed as intended (if any), and why
- It should include a cost-per-participant.

Detailed guidance for full-scale RCTs can be found in our Evaluation Guidance.

### Full-scale quasi-experimental evaluation (QED)

Studies that are not suitable for a full-scale RCT are considered for a full-scale QED, unless there is a consensus that this option carries a high risk.

A full-scale QED is similar to a full-scale RCT, with the exception that participants are not randomised. Priority is given to designs best suited to establish a strong causal estimate such as Regression Discontinuity Design, Difference-in-Differences, or Instrumental Variables.

Detailed guidance for full-scale QEDs can be found in our Evaluation Guidance.

### Pilot RCTs

Studies that are not suitable for a full-scale QED are considered for a pilot RCT, unless there is a consensus that this option carries a high risk.

In line with existing definitions, pilot RCTs are summative and formative in equal parts:

- The full-scale RCT, or parts of it, including the randomisation of participants, is conducted on a smaller scale to see if it can be done and what the results look like.
- Importantly, it involves trying out different delivery processes (such as different recruitment methods) and different evaluation processes (such as different ways of collecting outcome data). Rapid-cycle design and testing is strongly encouraged.
- They are expected to include mixed-methods research around the acceptability of the intervention and research design.

Detailed guidance for pilot RCTs can be found in our Evaluation Guidance.

### Pilot QEDs

Studies that are not suitable for a pilot RCT are considered for a pilot QED, unless there is a consensus that this option carries a high risk.

A pilot QED is similar to a pilot RCT, with the exception that participants are not randomised.

Detailed guidance for pilot QEDs can be found in our Evaluation Guidance.

### Feasibility studies

Studies that are not suitable for a pilot QED are considered for a feasibility study, unless there is a consensus that this option carries a high risk. Given the structure and sequences of our pipeline, we do not expect to fund many feasibility studies.

A feasibility study is primarily formative. Its aim is to ensure that the intervention and the evaluation design are ready for piloting. Thus, a feasibility study has two parts, one pertaining to the intervention, and one pertaining to the evaluation.

Each part is expected to combine theoretical and empirical evidence. Theoretical evidence is needed to interrogate and refine the theory of change underpinning the intervention (i.e. the mechanisms by which the intervention is expected to deliver the intended outcome) and the evaluation (i.e. the mechanisms by which the data is expected to yield credible evidence of impact and implementation). This evidence is likely to come from evidence reviews, workshops, and consultations with key stakeholders and experts.

Empirical evidence is needed to compare the effectiveness, feasibility and acceptability of different versions of the intervention (for example home visits vs. external meetings) and evaluation (for example, RCT vs. QED or primary vs. secondary outcome data collection). This can be obtained by means of rapid-cycle design and testing, observations, interviews, reviews of existing datasets, etc.

Detailed guidance for feasibility studies can be found in our Evaluation Guidance.

### Self-evaluation

Studies that are not suitable for a feasibility study, or happen outside of an open funding round can access the PINE portal to undertake a self-evaluation.

The aim of PINE is to turn promising practice into evaluable interventions and to improve the sector's demand or and readiness for research.

**Comparison of the types of evaluations**

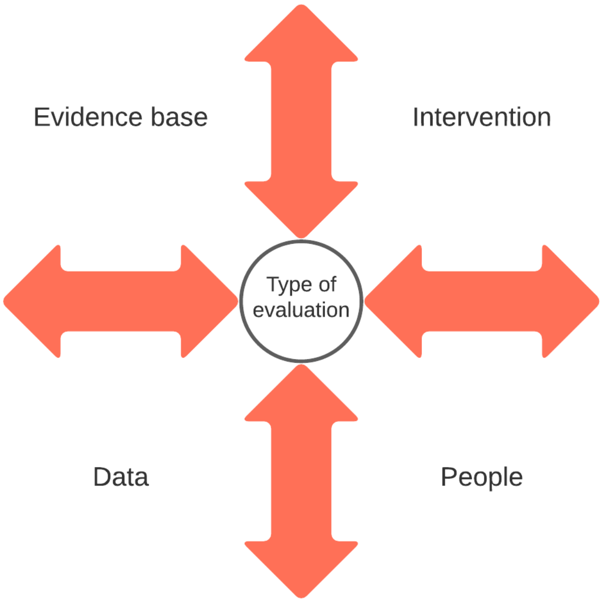
| | **Full-scale RCTs & QEDs** | **Pilot RCTs & QEDs** | **Feasibility Studies** | **Self-evaluation (PINE)** |
|---|---|---|---|---|
| Expected sample size* | Large | Medium | Small | Small |
| Emphasis on summative questions | Stronger | Medium | Weaker | Weaker |
| Emphasis on formative questions | Weaker | Medium | Stronger | Stronger |
| Focus of variation | Intervention A vs. Intervention B or BAU | Intervention A vs. Intervention B or BAU<br>And<br>Process A vs. Process B | Process A vs. Process B | No required variation but it can be included |
| Comparison of different research designs | No | No | Yes | No |
| Likely timescale | 12-18 months | 12-18 months | 6-18 months | None |
| Funding for delivery | Yes | Yes | Yes | No |
| Funding for independent evaluation | Yes | Yes | Yes | No |
| Support accessible outside of funding rounds | No | No | No | Yes |
| Report published on WWCSC website | Yes | Yes | Yes | No |
| WWCSC Data Archive | Yes | Yes | No | No |

(*) Based on power calculations

## 3. Risk factors

New programmes are first assessed against general eligibility criteria which are detailed in our calls for proposals, and against our overarching research principles.

In line with the process described in sections 1-2, programmes that are considered eligible for funding are first considered for a full-scale RCT unless there is a consensus that this type of study carries a high risk, using the factors below. If the programme is not suitable for a full-scale RCT, one or more new risk assessments are carried out, using the same risk factors, until the right type of evaluation is found.



Our assessments consider four risk *factors* (as opposed to objective *criteria*), which should be assessed separately in the interests of simplicity. We recognise that factors will at times be in tension. When this happens, we use the guiding principles described above as well as the available expertise to find a way forward.

**Evidence base**

We review the literature to assess the volume and relevance of existing evaluations. The decision to assign a programme to a type of evaluation can be conceived as a matrix.

| | | **Closeness of replication** | | |
|---|---|---|---|---|
| | | **Far** | **Moderate** | **Close** |
| **Volume of literature** | **Non-existing** | Self-evaluation or Feasibility study | | |
| | **Low** | Feasibility study | Pilot RCT/QED | Full-scale RCT/QED |
| | **Medium/ high** | Pilot RCT/QED | Full-scale RCT/QED | Evaluation likely to be redundant |

Some programmes have an evidence base that can be described as ‘thin and diluted’. In other words, their evidence base includes no or a few evaluations, and these studies are loosely related to the programme being considered. For example, the intervention is the same, but the population of beneficiaries and the evaluation designs are different. In this case, and irrespective of other factors, a small-scale evaluation (self-evaluation or feasibility study) might be the best option.

Conversely, some programmes have a large and rich evidence base. In other words, their evidence base includes a large number of studies that are directly relevant to the programme being considered (same intervention, same population, same research design, etc.). In this case, the evidence base is effectively saturated and any new evaluation may be redundant[1].

In reality, most programmes fall in the middle: their evidence base include a small number of studies that share some characteristics with the programme being considered (e.g. the intervention and the population are the same) but not all (e.g, the research design for the evaluation is different). In this case, a pilot or full-scale evaluation may be considered, depending on the closeness of replication, the risk level, the formalisation of the intervention, the acceptability of the design, etc.

Importantly, programmes are expected to enter the Evaluation Pipeline at least at the level of the best available evaluation. In other words, if the best available evaluation for a programme is a Pilot RCT (as defined by WWCSC), we would expect the WWCSC-funded evaluation to be at least a Pilot RCT, unless there is a consensus that another stage is preferable. Whenever possible, we will push to progress the evaluation to the next stage (in this case, full-scale RCT), in line with our scaling principle.

**Far vs. close replication: some examples**

Catch Up® Literacy is an example of project closely replicating a previous trial. It was commissioned following WWCSC's re-analysis of EEF trial data, which aimed to identify projects that have the potential to close the attainment gap between young people who have had a social worker and their peers. There are two main differences between the EEF trial and the WWCSC trial, namely: the target population (looked-after children in the WWCSC trial, disadvantaged children in the EEF trial), and the deliverers (foster/kinship carers in the WWCSC trial, teaching assistants in the EEF trial). Other trial parameters are identical, including the nature of the intervention (15-minute reading sessions, twice a week for Key Stage 2 pupils), the outcome (reading skills), the evaluation (RCT).

Kitbag is an example of far replication: WWCSC decided to fund a trial based on the findings from a qualitative study in 70 schools. However, this is also extensive literature on the effectiveness of trust-building activities on longer-term outcomes.

[1] A meta-analysis of this literature should be considered.

### Intervention

Any social work practice can be conceived as an ‘intervention’, that is, an external stimulus (such as counselling) intended to trigger a response (such as an improvement in well-being). Some of these interventions are well defined:

- They target people meeting pre-defined criteria, and exclude people not meeting them
- They are ‘protocolised’ or ‘manualised’, i.e. they are broken down into a sequence of simple activities or decisions, which are all documented
- They have clear and realistic objectives
- They are costed
- They can be compared
- They can be taught

Other interventions are more loosely defined. Well-defined interventions are well suited for full-scale evaluations or pilot evaluations, because they are easily distinguishable from other interventions. However, well-defined interventions tend also to be more studied. If there are too many evaluations of the same intervention, the evidence base is saturated, and a new evaluation may be redundant[2]. Conversely, well-defined but under-studied interventions are prime candidates for a full-scale or pilot evaluation.

When looking at an intervention, it is also important to consider likely benefits and harms:

- Interventions that have obvious or well-documented benefits, through previous RCTs for example, may not need to evaluated
- Interventions that we know are harmful should not be delivered in the first place
- Interventions with a balance of benefit and harm (equipoise) are prime candidates for an evaluation.

### People

Interventions have implementers, beneficiaries and evaluators. All groups are important in an evaluation, but for different reasons.

Beneficiaries matter in both quantitative and qualitative terms. From a purely quantitative perspective, beneficiaries give statistical power to an impact evaluation. Programmes that are delivered to large cohorts of beneficiaries should be considered for a full-scale evaluation. It is important to be very cautious when making this decision, as recruitment targets are often overoptimistic (optimism bias). Delivery partners should have evidence of demand/need and explicit recruitment assumptions. Conversely, programmes that are expected to be popular and oversubscribed are good candidates for RCTs with a waitlist.

From a qualitative perspective, beneficiaries can affect the quality of the data collected. Programmes that are delivered to reluctant beneficiaries are likely to result in higher attrition, lower quality data, or both. This point is particularly important in RCTs and in programmes

[2] This assumes that these evaluations are close replications, i.e. same population, same outcomes, same analytical methods, etc. Well-defined interventions evaluated with a wide range of methods would be good candidates for a large-scale evaluation.

where the benefit or harm is perceived to be high. In these instances, it may be wise to start with a pilot RCT or QED.

Implementers matter in the sense that they are gatekeepers, informants and data providers. A low level of engagement or insufficient readiness for a RCT is likely to make a full-scale RCT very risky. Two strategies can be considered in such a case: a full-scale QED or a pilot RCT testing different ways to maximise implementers' engagement in a rapid-cycle fashion.

When a programme and evaluation have ambitious recruitment and delivery targets, all parties are strongly advised to provide evidence of long-term demand and supply, and discuss the viability of their plans, much as a company would do.

When re-assigning a programme to a type of evaluation other than the one initially planned, it is important to ensure that the evaluation team still has the capacity and expertise to deliver. This recommendation applies in particular when an RCT becomes a QED.

Importantly, WWCSC will seek opportunities to generate evidence about minoritised groups and recognise that in order to achieve, we must work with implementers, beneficiaries and evaluators from diverse backgrounds. Equality, diversity and inclusion will be considered as a matter of course on every programme at the application stage and throughout.

### Data

Data collected for the sole purpose of an evaluation is called *primary data*. Data collected at a national or local level for other purposes, such as policy-making, project delivery and management, routine reporting or longitudinal studies (census data, programme monitoring data, exam results, etc.) is called *secondary data*. This includes, for example, attainment data for all children and care data for children in need.

When the outcome of interest is not routinely collected by delivery partners, primary data is sometimes the only option. It may generate data that is more valid, that is, more closely aligned with the objectives of the intervention (for example, the meaning of 'No Further Action' in care proceedings can be contentious). However, it is also much more expensive (as it needs to be collected), less reliable (the quality and timeliness of data may vary significantly from one organisation to another), more burdensome on participants (which means that the number of data collection points can only be limited), and often less representative (missing data is rarely random). When there is a doubt about the volume or quality of primary data that can be collected for an evaluation, it is usually safer to start with a pilot evaluation, or a feasibility study.

Secondary data has the opposite pros and cons. It is sometimes less valid (as it was designed and collected for a different purpose); however, it is considerably cheaper to use, more reliable and places no additional burden on beneficiaries and implementers. When valid, secondary data can also increase the precision of the impact estimate by providing a baseline (that is, the outcome pre-intervention) and by making it possible to conduct longitudinal analyses and quasi-experimental evaluations with high level of validity (such as difference-in-differences). Thus, evaluations using secondary data can be considered for full-scale evaluation, unless it presents other risks.

**Summary**

| | |
|---|---|
| **Factors expected to decrease the level of risk** | ● Programmes that are 'moderate' replications of a large number of studies or programmes that are close replications of a small number of studies<br>● Well-defined but under-studied programmes<br>● Programmes with equipoise<br>● Programmes targeting large cohorts of beneficiaries<br>● Programmes expected to be popular/oversubscribed<br>● Programmes where implementers are engaged and ready for trial<br>● Evaluation team capacity and expertise<br>● Programmes that use secondary outcome data |
| **Factors expected to increase the level of risk** | ● Programmes that are far replications of a small number of studies<br>● Loosely defined programmes<br>● Programmes without equipoise<br>● Programmes targeting small cohorts of beneficiaries<br>● Programmes expected to be less popular among beneficiaries<br>● Programmes where implementers are not engaged and not ready for trial<br>● Programmes that use primary or local outcome data |

## 4. Risk assessment output

A risk assessment will be undertaken by WWCSC for each new delivery project meeting certain criteria (such as those shortlisted to be paired with an evaluation in an open funding round). The assessment will be regularly updated and refined, until there is a consensus about the proposed evaluation. The final design is then published in the evaluation protocol.

Evaluators are encouraged to follow the same approach throughout the application and set-up stages.

**Risk assessment card**

A risk assessment card is essentially a document reviewing each key factor *independently*, that is, ignoring other risk factors. Where possible, it may be a good idea to assign each assessment to a different person, to reduce the risk of groupthink. A synthesis should be done at the end of the process, resulting in a risk level (see below).

| | |
|---|---|
| **Type of evaluation considered** | E.g. Full-scale RCT |
| **Evidence base** | Reasoning and conclusion (RAG rating) |
| **Intervention** | Reasoning and conclusion (RAG rating) |
| **People** | Reasoning and conclusion (RAG rating) |
| **Data** | Reasoning and conclusion (RAG rating) |
| **Other considerations** | Reasoning and conclusion (RAG rating) |
| **Risk level** | RAG rating |

The depth of the risk assessment should be proportionate to the level of agreement between parties about the likelihood that an evaluation will not meet its objectives. In cases where a consensus is rapidly obtained, a light-touch risk assessment is appropriate. In cases where there is no obvious consensus, a more thorough risk assessment is needed. A comparison of two designs is required when the risk level is Amber. Note that all decisions should be evidenced and documented.

**Risk levels and decisions**

| Level | Meaning | Decision |
|---|---|---|
| **Red** | There is a consensus that the design under consideration carries a high risk. | Move on to the next best design/stage |
| **Amber** | There is a consensus that the design under consideration carries a medium risk; or there is no consensus about the feasibility of this design | Assess the risks of the next best evaluation design and compare the two designs being considered to identify the most appropriate one. |
| **Green** | There is a consensus that the design under consideration carries a low risk | Implement the design; terminate the risk assessment. |

## 5. Going up the Evaluation Pipeline

Our vision is that all programmes and evaluations fulfilling certain conditions should be considered for scaling up, as defined in section 1. For example, Feasibility Studies fulfilling these conditions should be considered for a Pilot RCT or QED. We will consider both the findings of the evaluation and the context to make this decision.

Outcomes

- Evidence of impact
- Evidence of promise
- Evidence of participant engagement

Processes

- The intervention is evaluable
- The proposed evaluation design is low-risk

Context

- Availability of funding at the time
- Alignment with WWCSC's priorities at the time

We aim to publish all evaluations, regardless of the findings. WWCSC will publish a short response to each new evaluation, indicating whether the project is going to be scaled up and why.