WFP EVALUATION

World Food Programme

SAVING LIVES
CHAN GING LIVES

Technical note

Quality standards for impact evaluations

Introduction

1. Impact evaluation is a methodology and statistical tool for measuring causal attribution through counterfactual thinking. Impact evaluation aims to provide a quantitative measurement for the causal effect of a defined cause (for example, intervention, programme, or policy) on one or several outcomes.
2. The WFP Evaluation Policy 2022 defines impact evaluations as measuring changes in development outcomes of interest for a target population that can be attributed to a specific programme or policy through a credible counterfactual.
3. This document provides quality standards on the process and content for impact evaluations in WFP. Quality standards for impact evaluations vary somewhat between scientific disciplines but have common minimum requirements throughout. Within a counterfactual framework, impact evaluation needs to provide, as a minimum, the following aspects:
    - **Numerical estimates of the magnitude of the causal effect** (i.e. a quantified estimate in a well-defined metric);
    - **A numerical estimate of the statistical imprecision of the estimated effect** (for example, 95% confidence intervals);
    - A **factual comparison group** (i.e. subjects not affected by the same cause).
4. While the above considerations are minimum requirements for an impact evaluation, *high-quality* impact evaluations require further considerations, which can be categorised under three key areas: technical requirements, use of the evidence, and process.
5. Technical requirements need to be satisfied to ensure that the evidence generated can contribute to global knowledge on what works best to achieve SDGs (Objective 1 in WFP's Impact evaluation strategy). At the same time, evidence use and process requirements need to be satisfied to deliver operationally relevant and useful impact evaluations (Objective 2) and maximise the responsiveness of impact evaluations in rapidly evolving contexts (Objective 3).

July 2024

Box 1: Definitions: impact and impact evaluation

Impact evaluation does not aim to find the causes of observed outcomes but rather focuses on the impact of some well-defined cause. Impact evaluation, therefore, is less ambitious than other evaluation approaches but usually can provide more rigorous evidence as it is based on fewer assumptions and more stringent designs.

Impact Evaluation does **not** refer to temporal changes (i.e. changes in outcomes after the policy/programme as compared to the status quo before the programme) but instead refers to counterfactual changes, i.e. an observed situation as compared to what it would have been like without the policy or programme.

One may note that the meaning of the word "impact evaluation" differs from the word impact used in the DAC criteria. More specifically, according to the DAC criteria, the definition of impact is: "*The extent to which the intervention has generated or is expected to generate significant positive or negative, intended or unintended, higher-level effects*". This differs from the definition of impact in impact evaluation.

The table below summarises the difference between Impact Evaluation and Impact as defined by the DAC criteria.

Impact Evaluation	Impact according to DAC criteria
- Measuring the changes in outcomes of interest for a target population that can be attributed to a specific programme/policy through credible counterfactual - Short, medium, long term changes (that is, any point in time in the Theory of Change - Causal Attribution - Counterfactual thinking: difference of potential outcomes - Quantitative with well-defined statistical framework (complemented with qualitative)	- The extent to which the intervention has generated or is expected to generate significant positive or negative, intended or unintended, higher-level effects - Longer term effects - Indirect, secondary and potential consequences of the interventions - Examines the holistic and enduring changes in systems or norms - Quantitative and/or qualitative

Technical requirements

6. This section discusses the technical requirements to provide rigorous evidence for a high-quality impact evaluation from a methodological perspective. While the introduction already provided the minimum requirements for an impact evaluation, *high-quality* impact evaluations further require:
    - well-defined **statistical framework**;
    - appropriate **identification strategy**;
    - precise **measurements** of key outcomes of interest;
    - sufficiently large **sample size** and a clearly defined **sampling strategy**;
    - **data analysis** is based on a pre-analysis plan (PAP) and ex-ante planned evaluation designs.
7. This section will present and expand on each of these points.

Statistical framework

8. Impact evaluation should include the definition and clearly discuss the following components for the statistical framework:
    - Definition of a **population** consisting of units (e.g. individuals, households, schools, villages, etc.) that could be affected by the treatment (e.g. receiving a food supplement). For each unit, it shall be assumed that it could be affected by the treatment as well as not be affected by the treatment (i.e. a household may receive a food supplement or may not). These two different states are referred to as the treatment-state and non-treatment-state.

- Definition of **potential outcome states** Y1 and Y0, where Y1 is the outcome for a unit in the hypothetical state of having been treated (i.e. affected by the cause) and Y0 is the hypothetical outcome for the same unit in the alternative state of non-treatment.
- Reference and definition of a **causal parameter**, such as the Average Treatment Effect (ATE), Average Treatment Effect on the Treated (ATT), Local Average Treatment Effect (LATE), Intention to Treat Effect (ITT), or Quantile Treatment Effect (QTE)[1].
- Definition of the **treatment participation indicator** (D_i) which refers to the factual treatment state. D_i is either one or zero.
- Discussion of the assumption of the **absence of treatment spill-overs** or externalities or general equilibrium effects and its plausibility; also referred to as stable unit treatment value assumption (SUTVA), which says that the potential outcomes of unit *i* do not depend on the treatment realisation of other units.[2]
- Discussion on how units have been **allocated to the treatment**. In impact evaluations, it should be conceivable that for each unit *i* it might or might not be treated. Hence, the population of units is defined first, and discussion on the assignment process to treatment and control is defined afterwards.

Identification strategy

9. Impact evaluations are defined as a comparison of observed outcomes of beneficiaries to their hypothetical outcomes in the counterfactual state of non-treatment. Since counterfactual outcomes are obviously unobservable, they need to be estimated from the observed outcomes of the control group. Various experimental and quasi-experimental methods exist to estimate counterfactual outcomes, which rely on different assumptions.
10. The following designs are permissible:
    - ***Randomised Controlled Trials (RCT);***
    - ***Regression Discontinuity Design (RDD);***
    - ***Quasi-experimental designs;***
    - ***Natural experiments.***
11. In the section below, we will discuss the basic idea and key assumptions for each of these designs.

Randomised Controlled Trials (RCT)

12. Randomised Controlled Trials are considered the gold standard and are advised for impact evaluations. RCTs are also often implemented in the form of Randomised Encouragement Designs or Randomised Offer Designs, where the intervention or programme is offered to some individuals, households, or villages, who may refuse to take the offer.[3] In these designs, the randomisation indicator refers to the offer and not the uptake.

Regression Discontinuity Designs (RDD)

13. The Regression Discontinuity Design is applicable when interventions are allocated on a needs-basis or means-tested basis and limited programme budgets imply that many individuals/households/schools/villages above

[1] ATE measures the average marginal effect in the population, ATT measures the effects only among those receiving the intervention. ITT measures the effects among those that were offered to participate in the programme, even if those indivuals may have not ended up taking up the treatment. LATE measures the effect of the treatment on those who comply with the offer, but are not treated otherwise. QTE measures the effect at different points of an outcomes distribution.

[2] Absence of treatment spill-overs, externalities, SUTVA: The impact evaluation design shall also ensure the absence of treatment spill-overs, i.e. that the control group is not affected by the treatment offered to the treatment group. In cases of likely geographical spill-overs, the design should ensure a reasonable minimum geographical distance between treatment and control groups. For example, if inhabitants of neighbouring villages frequently interact, one may consider a minimum distance between any of the villages included.

[3] For example, offer of a nutrition supplement to children who might take the offer and eat the supplement or may refuse.

the threshold cannot be served. The RDD design has a strong disadvantage compared to RCTs as it provides impact estimates only for the group at the margin and thus provides information only for a small subset of the population. The RDD can often be considered a local randomisation design. Since the (local) randomisation is not under the control of the evaluation team, though, it is prone to possibly selective manipulation. To be considered valid, various tests for design validity should be done, including density tests.

Quasi-experimental designs

14. Apart from these two designs, which implicitly or explicitly are based on randomised treatment allocation, numerous other quasi-experimental methods exist, which rely on different assumptions. Only methods that implement and pass "placebo-treatment" tests or "falsification tests" shall be permissible for high-quality impact evaluations[4].

15. These designs and assumptions include:

    - **propensity score matching** with pre-programme tests;
    - **differences-in-differences** (or other synthetic control methods) with pre-programme alignments;
    - **instrumental variable** with overidentification tests and a credible argumentation for the absence of a direct effect

16. All these methods are more data-demanding than RCT and RDD, as they also require data for at least two pre-intervention periods before the intervention begins to allow for placebo-treatment tests[5]. Quasi-experimental designs that do not permit placebo-treatment tests (as well as those that do not pass the tests with the actual data) shall not be considered high-quality impact evaluations.

17. Instrumental variable approaches may also be considered in certain situations if a credible case for its validity can be provided. However, in practice is extremely difficult to find a good instrument that is correlated with programme participation but uncorrelated with the outcomes. Such instruments are likely to emerge from detailed knowledge of the mechanisms and contextual factors. In addition, an overidentification test shall be conducted, and in case of rejection, the approach shall not be considered as high quality anymore. Finally, since the statistical power of IV overidentification tests is low, one would request a convincing narrative for the necessary instrumental variable exclusion restriction.

Natural experiments

18. Finally, natural experiments may also be considered when a credible case can be made that the observed randomness is indeed completely random. Typical examples are natural disasters where the geography of occurrence could not be anticipated. This approach, however, does not seem to naturally fit with impact evaluations for WFP interventions.

Measurement

19. Indicators, measurement tools, and data collection processes are of key importance for the quality of impact evaluations. It is critical that high-quality impact evaluations make use of precise measures of key outcomes of interest.

20. The programme's Theory of Change (ToC) is typically the starting point for defining which indicators should be measured as part of the impact evaluation. High-quality impact evaluations shall describe in the ToC how the intervention is expected to change outcomes and critically discuss the logic and assumptions behind it. While in

[4] Note that placebo-treatment tests do not directly test the assumptions underlying the quasi-experimental evaluations; it is commonly considered likely, though, that a failure of the placebo-tests also reduces the plausibility of the main identification assumptions.

[5] Placebo tests are now conducted by pretending as if the programme/intervention had started two periods earlier and applying the estimator with this modified start date. An estimate (statistically) different from zero indicates that the assumptions underlying the quasi-experimental method are likely wrong. For difference-in-difference designs, this is referred to as "test of the parallel trend" or "test of the common trend" assumption. For matching or selection-on-observables designs this is referred to as "pre-programme test".

most programme documents, the ToC is outlined as a sequence of events and activities happening over time, with the thinking based on a mental before-after comparison for the beneficiary group. In impact evaluations, however, the ToC also has to embed and discuss a control group perspective, including some reflection about what would happen in the control/comparison group[6].

21. While the primary outcomes of interest for a programme's ToC are medium-term changes of substantial interest, it is also advised to have a few endpoints early in the results chain in the impact evaluation that can be measured precisely.[7]

22. Measurement tools -such as individual or household questionnaires - are of key importance for capturing indicators. Many of the outcome variables, such as empowerment, household decision-making, and violence, are often difficult to measure. Arguably, an impact evaluation is not the right place to test new measurement tools. It is advised to only use measurement tools that have been tested in other settings before. Survey design is an important research field on its own and is used for questionnaire design. For an impact evaluation, though, only validated tools should be used. Key outcome indicators and other outcome measurements should have been validated in prior research. Key outcome indicators should also be aligned with indicators used in other countries to permit cross-country comparability. In addition, survey questionnaires should be piloted and field-validated in the local context and local languages prior to data collection. Finally, the evaluation should also consider the precision of the measurement (i.e. the amount of measurement error or the signal-to-noise ratio).

23. Impact evaluation should place particular attention on how data are collected in the field. As much of the information collected is time-sensitive, data collection in the treatment and control groups should be conducted simultaneously. Similarly, data on programme participation should be collected in both treatment and control groups using the same tools. Data collection shall be done by independent data collection teams (i.e. not wearing WFP uniforms or using WFP cars) in order to avoid biased responses (courtesy bias). Finally, household surveys should have clear protocols on whom and how to interview within the household.

Sampling and statistical power

24. Inferences should be made from observations derived from an adequate sampling strategy and with a sufficiently large sample size that provides reasonable statistical power.

25. Sampling should be based on (i) random sampling from a clearly defined sampling frame (ii) a positive sampling probability for every unit in a defined population and (iii) a known sampling probability, which permits the use of weights in the econometric analysis.

26. The sampling process should also describe respondent selection and protocols for multi-stage sampling (if applied). For example, it should explain the protocol for respondent selection in the case of data collection from individuals within households.

27. Evaluation reports shall also provide information on non-response rates for treatment and control groups. Non-response rates should be similar for both groups, and differences of more than 10 percentage points raise concerns about the comparability of the groups. Non-response rates higher than 20% in any group will also raise questions about the representatives of the sample.

28. Impact evaluations shall perform power calculations to identify an adequate sample size. If impact evaluations are based on small sample sizes, the estimated effects may turn out to be statistically insignificant when in reality, the effect was present[8].

---

[6] For example if other forces are at play that also change outcomes in the absence of the intervention or if individuals may implement or access similar or other treatments.

[7] Consider for example an intervention that provides trainings (e.g. life skills, financial literacy, household gardens) aiming at increasing household incomes. Estimates of income effects may be statistically insignificant (if power calculations were too optimistic), which renders the interpretation of the impact evaluation difficult since the findings may indicate that the trainings are ineffective or may be due to too small sample size. Indicators such as training participation and knowledge can indicate if the intervention already failed at the output level and/or at the immediate outcome level.

[8] Those findings are often interpreted as indicating that an intervention is not effective, even though a too small sample size could be the main reason for this result. There are several reasons for why many impact evaluations are underpowered, for example, a

29. Power calculations should be based on a literature review of impact evaluations of comparable programme interventions to inform what would be a reasonable effect size. Alternatively, if programmes are very innovative and no literature review is possible, a Minimum Detectable Effect Size (MDF) not greater than 0.1 should be used unless it is adequately explained and justified.

30. Reports should provide reference to the statistical software package used for conducting power calculations. Conventional values are alpha 5%, beta 80%, and two-sided 5% tests. Power calculations shall also incorporate at least 10% potential non-response at the endline and various ICC scenarios obtained from already existing survey data for clustered designs. A discussion about implementation fidelity, treatment crossover, and complier rate is also required[9].

31. Finally, power calculations should also explicitly account for subgroups analysis expected in the evaluation (for example, gender or treatment arms).

Data analysis

32. All impact evaluations should outline the analysis the evaluation intends to conduct in a Pre-analysis Plan (PaP) or Inception Report (IR). A PaP should be prepared and published prior to endline data collection. The PaP should provide information on evaluation design, econometric methods and other considerations. The PaP should also state the intended subgroups for which effect heterogeneity shall be considered, indicate which data was available at the time of the writing, and indicate how missing data shall be handled in the econometric analysis.

33. The order in which analysis is also conducted matters. This section also presents what should be the sequence of econometric and statistical analysis. First, (a) econometric analysis without the outcome data Y, then (b) econometric analysis according to PAP, followed by (c) econometric analysis of effects on Y with ex-post modifications (if applicable) and finally (d) any further econometric analysis involving the outcome variable Y.

34. In the early stages of the analysis, one should implement analysis of (1) data collection quality, handling of missing data; (2) implementation data, MIS data, monitoring data from various sources; (3) implementation fidelity and treatment crossover in the control group, without examining the outcome data Y.

35. Subsequently, one would conduct the econometric analysis according to the pre-analysis plan. The econometric analysis following the PAP should be reported at least in the appendix. Even if the results might be of limited use due, for example, to operational disruptions or changes in implementation.

36. In many impact evaluations, the main econometric analysis is often modified from the original design. This might be due to unexpected events such as conflict, disasters, and natural calamities, or it might be due to substantial deviations from the programme implementation plan or control group contamination. In such scenarios, one would like to modify the analysis plan in order to tailor the analysis to the actual implementation and existing data. The report should always mention very explicitly which analysis follows the ex-ante PAP and which analysis modification has been developed ex-post.

---

limited budget that does not permit a sufficiently large data collection. Another reason could be overly optimistic performance targets, which assume a large impact (usually the target) and lead to underpowered evaluation studies if the target was too optimistic. A further reason could be optimistic assumptions with respect to implementation fidelity, which later may turn out to be lower because of either operational delays and/or availability of same or similar treatment in the control group.

[9] Statistical power depends on the degree to which the treatment group actually benefits from the treatment *and* the degree to which the control group has access to the same or a similar treatment. In the ideal design, 100% of the treatment group are affected by the treatment and 0% of the control group. In practice this is rarely the case and complier rates are less than one. Operational delays in implementation or partial implementation may imply that the treatment group may only partly benefit from the programme (during the evaluation period). In addition, the control group may have access to similar programmes, e.g. if a school feeding programme is implemented by WFP in some (possibly randomly selected) districts and it happened that a similar programme is also available in the control areas, e.g. by other NGOs, governmental programmes etc. This needs to be considered in the power calculations.

37. Finally, many impact evaluations ask questions beyond estimating the treatment effects. Additional analyses could be performed [10] as they often provide very important insights. However, it needs to be acknowledged that they often rest on stronger assumptions and are often developed after having seen the main estimates. Therefore, analysis conducted beyond the main impact questions should clearly be presented separately in the report.

Evidence use

38. WFP's impact evaluation strategy identifies the need to contribute to global knowledge as well as the need to test programme theories and learn what works best, how and for whom. This means that WFP high-quality impact evaluations need to be able to serve both a global scientific audience as well as a programme and policy audience.

39. This section will focus on the requirements to ensure that impact evaluation evidence is useful and relevant simultaneously to a global scientific and a programme/policy audience.

40. **Frame the study within already existing evidence**: Impact evaluations should provide the relevant and most updated literature review on the topic of enquiry. It should explain how the study is framed within the existing evidence and how it contributes to it.

41. **Conduct impact evaluation after a prototype or a pilot phase**: Statistically powered impact evaluations usually require large sample sizes and therewith imply substantial costs. Therefore, rigorous impact evaluations would be most useful after early implementation flaws, which are typical during the initial implementation stages, have been eliminated. Pilots will also provide valuable information for power calculations as well as programme process evaluations.

42. **Multiple treatment arms and effect heterogeneity analysis**: While the standard simple impact evaluation model considers a single intervention compared to a control group, for implementation research and operational learning, it is important to explore how different programme components generate differential impacts and affect different parts of the population differently. Ideally, impact evaluations should also considers several treatment arms and/or subgroup heterogeneity analysis.

43. **Additional qualitative data collection and analysis**: Even though impact evaluations are primarily quantitative, additional qualitative research methods can provide crucial contextual knowledge and important insights about possible explanatory mechanisms that would enrich the interpretation of the quantitative findings.

44. **Academic novelty**: For publications targeted to a global scientific audience academic novelty is an additional requirement to be publishable in a reputed scientific journal. Replication studies or evaluations of previous programmes in different contexts will often be considered as being of limited academic novelty. Nevertheless, such evaluation studies will be important ingredients for a meta-analysis, which itself would be of substantial interest. For academic publications in economic journals, a single impact evaluation study needs to build upon the existing knowledge base and provide incremental steps towards learning, for example by learning about the causal mechanisms and the differential contribution of various programme ingredients[11].

45. **Adequate timing**: For practical and logistical reasons, many impact evaluations estimate only short-term effects (for example, one or two years). However, the impact evaluation needs to allow enough time to be reasonable for the outcomes under analysis to materialise. Certain outcomes such as nutrition, empowerment and resilience take time to emerge. Therefore, it is important that the evaluation clarifies what it is expected to

[10] Additional analysis includes mediation analysis, spatial econometric models, network models, structural econometric models, behavioural choices, multinomial models with endogenous regressors and others. Generally, one would expect the main analyses to be nonparametrically identified, whereas supplementary analysis could exploit parametric modelling approaches. For example, quantile regressions, probit models, fixed-effect regressions, regression models with interaction terms interpreted as diff-in-diff.

[11] Public programmes often offer a bundle of interventions simultaneously as it is presumed that holistic interventions are needed to tackle the underlying problems. Combining all interventions simultaneously in one design, though, does not allow to learn about the effects of each single piece or the complementary or substitutive effects of their combination. Multiple treatment-arms trials would be needed to disentangle the separate causal contributions of each component.

observe in the timeframe of the evaluation. Impact evaluations that are able to provide insights on medium- or long-term effects are preferable in order to learn if initial effects are sustained.

46. **Report findings:** The readability of reports is crucial for ensuring that evidence is accessible and useful. It should be noted that reports are not the only dissemination product, as blogs, briefs, presentations, and academic journal articles might also be considered and used for disseminating the findings. However, they will all make reference to the relevant report as their main reference point.

47. The report consists of the main body and detailed technical appendices. While the appendices shall be rich in detail with respect to evaluation design and subsequent modifications, the main report should be largely non-technical with cross-references to the respective appendices where technical details are to be provided. A short summary should also be provided, including the main findings and considerations. The report should provide clear information on the question under analysis and how it relates with the already existing evidence; a clear description of the interventions' characteristics and its context; the evaluation design including sampling strategies, data collection tools, and additional qualitative methods; analysis, interpretation of the findings and considerations. The impact evaluation reports should clearly indicate which statistical analysis and regressions had been planned ex-ante (i.e. in the pre-analysis plan) and which regression analyses have been devised only after the outcome data became available. The analysis and interpretations of findings should make a clear distinction between the causal effects, the mechanisms, and the suggestive supporting evidence. Key estimates should be reported through visual aids such as charts, reporting the estimated effects and their statistical imprecision (for example, 95% confidence intervals)[12]. Finally, estimates to be presented to permit meta-analysis, such as effect size, standard error, control group mean, and sample size (in each treatment arm).

Process

48. In addition to technical and usefulness requirements presented in the previous section, there are important procedural aspects that need to be satisfied to ensure that generating evidence is providing value in itself and is not considered a burden for the beneficiaries and programme teams.

49. Impact evaluations need to be incorporated within ongoing or planned operations. There might be some potential trade-off in priorities between global knowledge production through scientific dissemination and ensuring that the impact evaluation process is aligned with programmes' needs. Reviewers from scientific academic journals will judge the merits of the research project, with potentially limited attention on procedural requirements or expectations. Nonetheless, attention to how the evidence is generated is crucial to ensure sustained demand and use within an organisation such as WFP.

50. This section presents important procedural aspects for high-quality impact evaluations.

51. **The evaluation meaningfully engaged stakeholders from the beginning and throughout the process**: Rigorous impact evaluations, in particular if Randomized Controlled Trials, require synchronisation and tight collaboration between the programme implementation team and the evaluation team. It is therefore crucial for the programme team to recognise the value and importance of the evaluation and the implications when deviating from the proposed design.

52. Stakeholders, however, can go beyond the programme teams and might, for example, include national governments or multinational organisations with interest in the topic. A detailed stakeholder map analysis might help identify such groups based on a common learning agenda and shared interest and desire to answer relevant questions.

53. **The evaluation meaningfully engaged expertise and local researchers in a mutually advantageous exchange**: At a minimum one would like to incorporate knowledgeable local researchers in at least three domains: For questionnaire design and adaptation of questionnaire manuals and survey items to local contexts, customs and wording. In addition, for those impact evaluations which include qualitative research components,

[12] Visual reporting of CI shall also help readers to understand that smaller sample sizes may lead to statistically insignificant effects. E.g. separate analysis by gender splits the sample size in half and thereby increases imprecision as compared to full sample analysis.

knowledgeable local experts would be required for interview guidelines and protocols. Finally, local expert knowledge should also be consulted for the interpretation of the impact evaluation results.

54. **Independence and transparency of the evaluation team**: While impact evaluations require a strong collaboration between the evaluation team and programme implementation team, the evaluation team needs to have complete independence over the findings. Moreover, researchers and evaluators involved in the study should have their incentives disconnected from the findings of the evaluation. This also applies to WFP staff, where evaluation team members should come from an independent evaluation unit with clear and distinct career paths and career progression incentives that are different from the programme's performance. Any potential concerns about a potential conflict of interest (if any) of the evaluation team shall be transparently disclosed.

55. **Impact evaluation should include learning aspects and go beyond the accountability purpose**: Even though the impact evaluation team must maintain independence, it nevertheless needs to absorb and incorporate detailed knowledge about the programme, its implementation and country operational plans. The process of engagement with the operations team in the country is also important for ensuring a cooperative working atmosphere and for the sustainability and use of impact evaluations. Such a cooperative working atmosphere is more likely to be maintained if the impact evaluation goes beyond a pure accountability purpose and accountability is not seen as the primary purpose.

56. **Evaluation obtained ethical clearance from an Institutional Review Board (IRB)**: Obtaining ethical clearance is a critical component in any impact evaluation. This is, first and foremost, to ensure that none of the practices in the study might create any harm or risk to the people we work with. Second, there are reputational considerations and risks that might affect WFP's reputation if evaluations and studies are not conducted ethically. Finally, in order to publish in an academic journal, approval from a recognised IRB is often a requirement. It is important to keep in mind that the ethical clearance should refer to the research/evaluation part and not to the programme itself. The programme itself should have undergone ethical considerations beforehand during its planning stage.

57. **Adhere to data protection rules and regulations**: Impact evaluations rely on collecting, using and analysing large quantities of data, including personally identifiable information. This poses both ethical and juridical challenges. Therefore, it is important that the evaluation has clear protocols and procedures for dealing with sensible and personal data that are in line with the most updated rules and regulations.

58. **Data is made publicly available**: Once the study is completed, anonymised data and econometric programming code are made publicly available to other researchers. This serves a dual purpose. First, it treats data as a public good, supporting and incentivising other researcher to explore more questions and broader the evidence base. Second, it serves as transparency mechanism, allowing verification and replication of the published studies.

59. **The evaluation is explicitly considering equity and inclusion aspects, such as gender, people with disabilities, minorities, in line with WFP policies**: Impact evaluations shall go beyond estimating average effects and also examine impacts on inequality (e.g. via quantile treatment effects) or effects on different subpopulations, including vulnerable populations.

Impact evaluation feasibility checklist

1. Evidence needs and use: Would the proposed impact evaluation contribute to answering global and local evidence needs?	What questions does the proposed impact evaluation answer and how would it contribute to existing literature on the topics examined? Does the impact evaluation plan to examine cost-effectiveness?
2. Design: Is it possible to identify a credible counterfactual?	Please explain for which activities / components / interventions it would be possible to identify a credible counterfactual and how. Identification strategies can include experimental and quasi-experimental designs. All quasi-experimental designs require data from both the internvention and comparison groups for at least two pre-intervention periods before the intervention begins, to allow for placebo-treatment tests.
3. Programme duration: Is it realistic to expect to observe change within the proposed timeframe of the study?	For how long can intervention and comparison groups be maintained? Is it realistic to expect to observe change within this timeframe?
4. Sample size: Is it possible to identify a sufficient sample size?	Will the intervention and associated source of variation include enough units of observations to have statistical power? How many households / villages / schools will the evaluation include?
5. Implementation experience: Is there enough implementation experience to allow a stable programme implementation?	How familiar is the programme team with the activities implemented and the areas of intervention? Is the country office directly involved in the implementation of the activities or is this implemented through partners?
6. Budget: are there enough resources to allow all the relevant costs (including data collection and technical assistance)?	Are there enough resources to cover all the costs? This should include technical assistance as well as data collection costs. Data collection typically vary from country to country, and the sample size may vary depending on the evaluation questions. Please frame the resource assessment based on these considerations.
7. Timeline: Is there enough time to allow for designing an impact evaluation before programme activities begin/change/scale-up?	When are the beneficiaries expected to be identified and activities expected to start? Enough time should be allowed to define the evaluation questions and design, develop tools, register the study, obtain IRB approval, identify eligibility criteria, select units, and conduct a baseline.

Key personnel

CENTER FOR EVALUATION AND DEVELOPMENT (C4ED)

Markus Frölich	Full Professor of Econometrics at the University of Mannheim in Germany

OFFICE OF EVALUATION

Anne-Claire Luzot	Director of Evaluation
Jonas Heirman	Senior Evaluation Officer (Head of Impact Evaluation Unit)
Simone Lombardini	Evaluation Officer (Impact Evaluation)

Photo credits

WFP/ Samantha Reinders