NOVEMBER 2022

CEDIL Methods Brief 9

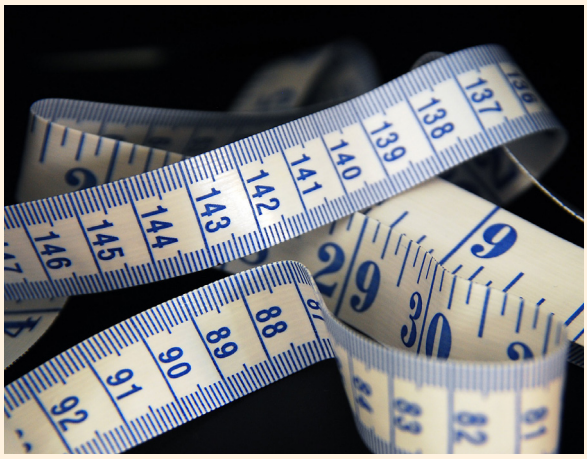
Innovations in measurement and the evaluation of human behaviour

Centre of Excellence for Development Impact and Learning

Box 1: Highlights

If policymakers or researchers aim to improve an intervention, scale it up, or assess why it works and for whom, then focussing solely on direct outcomes is not enough. There are several ways in which evaluations can go beyond traditional outcome measurement and explore the underlying drivers of impact and the behaviours of beneficiaries and non-beneficiaries.

- Measurement goes hand in hand with developing a theory of change, and the two work best when they evolve jointly: new (and better) measurements could allow the consideration of more flexible and realistic theories, while theory informs the construction and design of new measures.
- Linking measurement to human behaviour and decision-making processes offers useful angles for identifying the underlying drivers of an intervention's impacts.
- Measurement innovation and improvements—including the construction of new measures—are an important endeavour requiring cross-disciplinary input. This endeavour should be promoted.



Introduction

In evaluation work, the direct outcomes of an intervention often receive the most attention. However, if policymakers or researchers have additional goals—like improving the intervention, scaling it up, or determining why it works and for whom—then focusing solely on direct outcomes is not enough.

However, moving beyond the measurement of outcomes alone presents four key challenges and complexities. First, measuring outcomes involves a high degree of uncertainty. For instance, the outcomes affected by an intervention may not be knowable ex ante or during implementation. Likewise, some outcomes are not directly observable, making them hard to measure. Moreover, sometimes the existing metrics for measuring a particular outcome are not appropriate in a given context.

Second, it can be challenging to interpret what has been measured. Is an observed impact small or large? Is the impact important for target groups? Often, it can be difficult to find an appropriate, well-measured, and meaningful benchmark to help interpret results.

Third, many interventions affect multiple outcomes. These can include outcomes that an intervention explicitly targets, complementary effects, or other outcomes that constitute side effects, which themselves may be desirable, neutral, or undesirable (to the point that they can offset the intervention's desirable effects).

Fourth, and perhaps most importantly, traditional evaluation and measurement approaches often cannot identify the underlying drivers of a given outcome, whether positive or negative. Myriad factors can influence an intervention's success (or failure), and these often have to do with subjective factors and human behaviour (e.g. perceptions, expectations, preferences, and social norms)—which can vary from group to group and from context to context. Such considerations are particularly important when seeking to deploy a successful intervention at scale.

Untangling these challenges and complexities requires ingenuity and innovation, including the establishment of new metrics and measurement tools. In this brief, we present a number of critical reflections on how evaluation can go beyond traditional outcome measurement in order to explore the underlying drivers of impact.

Box 2: Why we should avoid an exclusive focus on direct outcomes

Focusing solely on an intervention's direct outcomes can lead evaluators to overlook the intervention's wider, but nonetheless important, effects. Consider an intervention that aims to improve nutrition for poor children by providing breakfast or lunch in primary school. While a straightforward impact evaluation might determine that the intervention does improve nutrition levels for beneficiaries, it might overlook other effects. Parents of the beneficiaries, for instance, might be reducing the food they provide to the target children at home, perhaps to provide more nutrition to their other children.

Likewise, a straightforward evaluation of providing additional (but less qualified) teaching assistants to certain schools or nurseries might find sizeable positive impacts, but the extra resources might also trigger a reallocation of teachers' time, which could reduce or even cancel out the observed impacts.

It is therefore important to model and understand individuals' behaviours (and their drivers), including those who are not direct beneficiaries of an intervention but are involved indirectly. This endeavour involves developing what in many disciplines is identified as a *theory of change*.

The challenge of measuring human behaviour

All interventions—and evaluations of interventions—involve an implicit theory of change about human behaviour. In practice, the theory of change that informs the design and evaluation of policy interventions should include the outcomes of interest, the individual behaviours that shape those outcomes, and the drivers of those behaviours—as well as the measurement of all these factors.

Many models of individual behaviour, particularly in economics, apply a theory of change that is premised on an individual behaving rationally. This approach sees behaviour as the product of an individual's assessment of the likely gains and costs from one action over another, while taking into account individual constraints. In evaluating policies, therefore, according to this approach it is key to take into account that individuals respond to policy interventions in order to achieve certain goals. While this approach might seem restrictive, it can accommodate very nuanced and sophisticated models of individual behaviour. Such models can incorporate a wide range of factors influencing behaviour, including: the resources and information available to an individual; the markets they can access; their preference for various outcomes; their beliefs about which outcomes will result from certain actions (which can be inaccurate or biased); their expectations regarding events or the actions of others; and social norms. While some of these factors are directly observable, others—particularly subjective factors related to individuals' preferences, beliefs, and expectations—are not.

To overcome the lack of data on these subjective factors, researchers often make strong modelling assumptions. For example, without data on subjective expectations, researchers might assume that each individual will rationally and efficiently use the information available to them. Likewise, if they are lacking data on individual risk appetites, researchers might assume that everyone is equally risk averse. In ideal circumstances, the availability of appropriate data helps avoid strong and arbitrary assumptions. However, in reality, what data are available often determines what can be measured and what models can be used. Likewise, developing more flexible models of behaviour could enhance the measurement strategies used by researchers (and evaluators, in particular).

While theories of change about human behaviour may seem to be unconnected to efforts to evaluate and measure the impacts of an intervention, they are actually inextricably linked. Moreover, strengthening this link can enhance evaluation and measurement efforts. Doing so underscores the importance of carefully identifying and using the most appropriate techniques and measurement tools to assess an intervention—not only its outcomes, but also the underlying drivers.

Measuring outcomes and their underlying drivers

If an evaluation aims to go beyond just establishing what the impacts of an intervention on a set of outcomes are and also to examine how those outcomes arise, one can divide the variables to be measured into *outcome variables and environmental or mediator variables.* Outcome variables are of direct interest to the researcher or evaluator: for example, the set of variables that a given intervention wants to influence or affect. Environmental variables, by contrast, are drivers of behaviour: for example, variables that might help to identify the mechanisms that generate the ultimate results of the intervention. Box 3 gives some examples of why measuring mechanisms is of crucial importance to the success of an intervention, and the policy implications one can draw from such measurements. What variables get included in the two sets depends on many factors, including the type of intervention being evaluated, the outcome variables of interest, and the feasibility of different evaluation methods.

The nature of the variables that will be measured (whether they are outcomes or environmental variables) is not the only issue for measurement. In some situations, outcomes and environmental variables are directly observable and established measures exist. In other situations, variables are difficult to measure either because they are not directly observable (e.g. attitudinal change) or because existing measures are not appropriate (e.g. standard measures that were constructed many years ago on samples typically from Western, educated, industrialised, rich and democratic (WEIRD) countries that are less applicable to developing countries). More generally, the presence of measurement errors should be taken into account, as these are pervasive even when certain measures are available and deemed appropriate. As a perfect measure does not exist, it can be better to have two imprecise measures of the same variable than it is to have one more precise one, provided the measurement errors affecting the two measures are independent of each other. In such a case, one can use one variable as an instrument for the second or, more efficiently, embed them in a measurement system that can provide efficient estimates of the unobserved (latent) variables of interest, as described below.

Box 3: Measuring the outcomes and mechanisms of a stimulation intervention

Consider an evaluation of an intervention to improve cognitive development targeted at young children in rural towns in Colombia that consisted of weekly home visits to improve parental investment in their children. Using a clustered randomised controlled trial, Attanasio et al. (2014) showed that the intervention had an impact of 0.26 standard deviations on children's cognitive development. But what drove that result? Did the impact actually work through increased parental investments or through some other aspect of the weekly home visits?

Different answers to this question would have significantly different policy implications. However, establishing a *causal link* between parental investment and child development is difficult, given the range of factors linked to child development. To overcome this challenge, Attanasio et al. (2020) examined the underling components of parental investment (e.g. material and time) and evaluate the intervention's impact on these components. Using an instrumental variable method to isolate the causal links, they provided evidence that most of the intervention's impact could be explained by increases in parental investment. But this finding raises another question: why did the parents increase their investment?

This question would be impossible to answer without data on subjective factors—like parents' baseline beliefs about the process of child development and the usefulness of parental investment in this process. Attanasio et al. (2019) utilised surveys from the Colombia intervention to estimate such beliefs, finding that parents typically underestimate the value of parental investment, particularly when children have high initial levels of cognitive development. Likewise, they showed that perceptions about parental investment are predictive of actual investment: parents who underestimated the value of investment invested less. Thus, a key aspect of the intervention's success is that it changed parents' beliefs about the usefulness of parental investment. Critically, understanding this process at play behind the intervention's impact was only possible due to the availability of data on subjective factors, like parental beliefs.

Updating and adapting measurement approaches

In any evaluation study, there are a number of overarching considerations related to how to use measures and how to measure impacts. These include the following: a) recognising the presence of and taking into account measurement errors; b) assessing trade-offs and options in regard to measuring behaviour through stated or revealed choices; c) interpreting impact sizes; and d) ensuring that measures can be compared across subjects.

Recognising measurement errors explicitly

The classical measurement error problem concerns errors that are inherent in the measurement of a set of variables. It is well known that the presence of such errors prevents the measurement of the correlation between different variables and, even more critically, establishing causal links. Four approaches can help reduce measurement errors. i) Use two (or more) measurements of the same factor and use the combined average measure in estimation; averaging may reduce the combined measurement error. ii) Develop an explicit measurement system which accounts explicitly for the presence of measurement error; this requires identifying the parameters that link underlying latent (unobserved) factors to their measures (observed indicators). Common methods are the multiple indicators multiple causes (MIMIC) model, factor models, or item response theory. iii) Adopt strategies in data collection to ensure that measurement errors relating to the variables of interest are independent among them: for instance, assigning different enumerators to collect different measures on the same subjects. While potentially costly, such measures can be very effective and can be cheaper than other alternatives. iv) Adopt design features that limit measurement errors and biases in the elicitation of responses, such as using appropriate experimental designs to minimise social desirability bias and demand effects.

Box 4: Measuring vindictive behaviour

Cavatorta, Zizzo, and Daoud (2022) were interested in evaluating whether exposure to violent conflict influences behaviour towards classmates, and potentially begets more violence. They looked at the case of Palestinian school children who cross military checkpoints (locations of frequent violent interactions) on a daily basis to go to school. The study compared retaliatory behaviour in groups of children who had an obligation to cross checkpoints on their way to school with that of other students at the same school who did not. Since self-reported behaviour on retaliation would be prone to bias, the authors used an interactive game between pairs of participants who played in turn. The game had clear and simple rules. Each player received an endowment: no aggressive action and no retaliation was the socially optimal choice for both players. However, the first mover could perform an 'aggressive' action towards the co-player by taking away some of the co-player's endowment. In the following turn, the co-player could decide whether or not to retaliate against the first mover. Retaliation in the game did not need to occur but, if it did, the second player's behaviour captured a marker for vindictive behaviour, which was comparable across participants. Every player was in the same situation and faced the same incentives and constraints. The directly observed behaviour could be correlated with (self-reported) stated choices or administrative data (e.g. school disciplinary actions). In this study, the authors found that youth who were obliged to regularly cross military checkpoints were significantly more likely to display retaliatory behaviour against their peers.

Measuring behaviour through stated or revealed choices

The measurement of individual behaviour can be divided into two categories: stated choices (i.e. self-reported behaviour in hypothetical situations) or revealed choices (i.e. what one actually does, measured by observing the behaviour directly). There are trade-offs in both approaches. While observing behaviour directly is often preferred, as it minimises different types of self-reporting bias, in many contexts this can be limiting or impractical. On the other hand, while measuring self-reported behaviour is highly common, particularly in surveys, it is not always feasible or appropriate: it may be context-dependent, as well as prone to misreporting and social desirability bias—considerations that are particularly important in relation to anti-social, stigmatised, or 'negative' behaviour.

The metric of impact sizes

Understanding 'what works' typically involves estimating the size of a treatment effect. A common method is to estimate the difference between control and treatment groups in terms of standard deviations. While such an approach makes sense when standard deviations reflect 'natural' variation in the outcome of interest, it can be misleading in other situations. For instance, an impact might look large in a very homogenous population but small when implemented in a very diverse population.

An alternative approach is to 'contextualise' the effect of a treatment, which can be done in three key ways. First, evaluators can compare the value of the treatment effects obtained in one context with known magnitudes from other studies or known values from meta-analyses. Second, they can compare the population under study to a representative sample from the same country or region comprising individuals from different (i.e. more or less 'well-off') backgrounds. As the example in Box 5 illustrates, this approach situates the target population relative to the overall population and contextualises the measured impact alongside broader population trends, indicating the extent to which an intervention remediates initial inequalities and helps disadvantaged groups to 'catch up'. Finally, evaluators can use a monetary metric: this requires monetising (to the extent possible) the benefits of an intervention and comparing them to the intervention's costs; however, it can be difficult to monetise the non-monetary effects of an intervention

Box 5: Measuring contextualised effect sizes

Many policy interventions seek to remedy social or economic inequalities by 'levelling up' beneficiaries' outcomes relative to those of less disadvantaged groups. Contextualising the effects is often important when measuring such interventions. For instance, Gertler et al. (2014) assessed the impact of an early childhood development intervention aimed at increasing later-in-life earnings for stunted children. As in traditional evaluation designs, they first compared the earnings of a group of stunted children receiving the intervention (treated group) with a group of stunted children not receiving the intervention (control group). To contextualise these effects, they then compared the treated group's earnings with those of *non-stunted* children. The study's results indicate that treated stunted children 'catch up' with initially better-off children, which means that the intervention was able to remediate initial inequalities. The study also provided a useful comparative estimate that could be used to evaluate the opportunity costs of different policy approaches for similar goals.

Ensuring measures are comparable across respondents

A frequent assumption when evaluating interventions is that subjective measures are comparable across people, such that meaningful summary statistics can be reasonably computed. However, this assumption is often unrealistic. Consider, for example, efforts to measure perceptions of an event's likelihood: the definition of qualitative categories such as 'unlikely', 'somewhat likely', 'likely', or 'very likely' can vary significantly across people. One person might define 'likely' as an eight in 10 chance, while another may define it as a five in 10 chance (see Wintle et al. 2019, Figure 5). This illustrates the problem of interpersonal comparability.

Measures exist that can alleviate this problem. One approach is to elicit perceptions of the likelihood of events. Because these measures are based on an objective metric (a probability), these measures represent improvements over traditional qualitative categorical measures and can be compared across subjects. However, eliciting probabilities comes with challenges. Lack of familiarity with the concept of probability can be a concern in some settings. Comprehension can be aided by the use of visual aids, careful explanation, and practice questions. Another concern relates to the possibility that perceived likelihoods are highly uncertain. Phrasing questions precisely (e.g. 'What is the minimum chance that the event will occur?') or seeking to elicit a respondent's entire probability distribution for a given question can help avoid these challenges. Eliciting the entire distribution is necessary when the variable of interest is the variance of the distribution, as opposed to the mean or mode. Box 6 describes a method for eliciting the entire distribution of income expectations.

Perceived likelihoods of events are inevitably subjective and the level of their accuracy is inevitably an issue. There is no guarantee that respondents will state the true subjective probability. Respondents may choose to distort their answers to rationalise past actions or report salient probability (e.g. 0.5 or 1) as a cognitive 'short-cut'. Incentivising respondents to tell the truth about their beliefs is one way to address this problem. A simple way to do this is to reward respondents when their guesses turn out to be empirically correct or approximately correct: that is, when an objective realisation or appropriate statistic is available (see Box 7 for an example).

Box 6: Eliciting the entire distribution of income expectations

When future outcomes are uncertain, individuals are likely to have a range of expectations about whether any specific outcome is likely to occur. People's expectations of their future income levels is a good example. Individuals will have a range of future income levels that they consider possible (from a minimum income to a maximum income), and a distribution of probabilities attached to each value in this range. The shape of this distribution is of interest because it reveals information about the expected income and its variance. Researchers have designed survey questions that help reveal these probability distributions. First, the perceived minimum and maximum future income level at a given point in the future (e.g. 12 months from now) must be established (this range constitutes the support of the distribution). Then the range can be divided into intervals, and the respondent is asked to indicate the probability that their future income will be less than a specified value. The intervals and the reported probabilities can then be used to infer the respondent's subjective expectations of their future income. This method has been implemented in several settings using different interview modes, including in-person interviews in developing countries (see Attanasio, 2009) and by telephone (see Dominitz, 2001; Cavatorta and Groom, 2020).

Box 7: Incentivising truth-telling responses

One's behaviour in a given situation can be driven by multiple factors, including what one expects others to do in response. In a further study of reciprocity among Palestinian school children who cross military checkpoints on their way to school (also discussed in Box 4), Cavatorta, Zizzo, and Daoud. (2022) were interested in measuring the participants' *expectations about* other people's behaviour, as measured using the simple gamified interactive task described in Box 4. Beliefs were elicited by asking 'How many participants in this room who play the role of first mover will take away some tokens from their co-player?'

Participants were rewarded if their answer was factually correct. This is a simple elicitation mechanism that is easy to understand (arguably easier than eliciting a subjective probability) and easy to implement in field studies. From a theoretical perspective, this method elicits the *mode* of the distribution over all possible empirical frequencies of an outcome. However, this method does not work for binary variables with no repeated draws (e.g. 'What is the likelihood that Italy will win the Euro World Cup?'), which require a probability estimate.

Improving and complementing existing measures

A standard practice in academia is to implement existing tests and measures that have been used for a long time. It is common, for instance, to use scoring mechanisms designed in a different context to that which is being studied and to employ measures that have been deemed reliable over time. While using existing practices has its advantages (as well-established measures and their properties are well known), such an approach is not necessarily efficient and can often lead to serious biases. When appropriate, it can be important to construct new measures, both in terms of the factors being measured and the tests being used.

For example, data reduction techniques like principal component analysis (PCA) are particularly useful for summarising a wide range of indicators. PCA is a statistical technique that constructs new variables as linear combinations or mixtures of the initial variables in a given data set, such that the new 'components' explain as much of the variation in the original data as possible. Often, the first factor (or the 'principal component') in PCA is the one that reflects that largest amount of variation from the original data. The second indicator is the one that loads most strongly onto the first factor (i.e. is more informative about it) and so on. For studies that have traditionally required respondents to answer a large and time-consuming battery of questions, PCA strategies have been adopted to optimise the collection of information (see Box 8 for an example).

Other novel measurement techniques come from the discipline of psychometrics. These include experimental designs and computerised interactive tasks that aim to reveal unobservable attitudes, such as stigmatising behaviour (see Cavatorta and Shukri, 2022).

Box 8: Summarising a wide range of indicators using PCA

The Bayles Scales of Infant and Toddler Development is a popular measure of child development, but it contains 91 survey questions that are typically asked one-by-one to a child's mother. Such a large number of questions limits the measure's use in environments where data collection faces resource or time constraints. To help address this challenge, PCA can be used to identify the survey's most informative elements. Utilising a sample of Indian toddlers whose cognitive ability was measured using the Bayes Scale, Attanasio et al. (2020) used PCA to identify the most informative elements, resulting in a set of 15 items. Used as a proxy for cognitive ability, the linear combination of these 15 items yielded a distribution equivalent to the one obtained using the entire set of 91 questions. Conducting the shorter questionnaire required approximately one-sixth of the time required to implement the full Bayles Scale. While many researchers may prefer to use the full questionnaire, the 15-item scale can be more practical (if not the only feasible approach) in many settings.

Conclusion

Better (and sometimes new) measurements are key to policy evaluations because many drivers of behaviour may not be directly or immediately observable. We argue that measurement and theory development go hand in hand, and they work best when they evolve jointly: new measurement informs more flexible and realistic theories of behaviour, while theory informs the construction and design of new measures.

The list of examples provided in this brief is far from exhaustive, yet it is clear that new developments are happening in this area and important progress is being made. The necessity of further improving measurement is ubiquitous in many disciplines and will require multi- and cross-disciplinary input. Large public initiatives like CEDIL can play an important role in this process, from facilitating research collaborations and interactions across disciplines to the standardisation of measures and the development of new measurement tools.

Endnotes

Attanasio, O.P. (2009) 'Expectations and perceptions in developing countries: their measurement and their use', *American Economic Review* 99(2), pp. 87–92.

Attanasio, O.P., Fernández, C., Fitzsimons, E.O., Grantham-McGregor, S.M., Meghir, C., and Rubio-Codina, M. (2014) 'Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial, *British Medical Journal*, 349.

Attanasio, O.P, Cunha, F., and Jervis, P. (2019) 'Subjective parental beliefs: Their measurement and role', *NBER Working Paper 26516*.

Attanasio, O.P., Cattan, S., Fitzsimons, E., Meghir, C. and Rubio-Codina, M. (2020) 'estimating the production function for human capital: results from a randomized control trial in Colombia', *American Economic Review* 110(1), pp. 48–85.

Cavatorta, E. and Groom, B. (2020) 'Does deterrence change preferences? Evidence from a natural experiment', *European Economic Review* 127, 103456.

Cavatorta, E. and Shukri, I. (2022) 'A novel method for measuring stigma in health: evidence from adolescents during COVID-19', *SSRN Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4034483

Cavatorta, E., Zizzo, D., and Daoud, Y. (2022) 'Conflict and reciprocity: A study with Palestinian youths', *Journal of Development Economics* (forthcoming).

Dominitz, J. (2001) 'Estimation of income expectations models using expectations and realization data', *Journal of Econometrics* 102(2), pp. 165–195.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., ... and Grantham-McGregor, S. (2014) 'Labor market returns to an early childhood stimulation intervention in Jamaica', *Science*, 344(6187), pp. 998-1001.

Wintle, B.C., Fraser, H., Wills, B.C., Nicholson, A.E., and Fidler, F. (2019) 'Verbal probabilities: Very likely to be somewhat more confusing than numbers', *PLoS* ONE 14(4), e0213522.

About this brief

This brief was prepared by Elisa Cavatorta and Orazio Attanasio. It draws on Attanasio, O. and Cavatorta, E. 2022. Evaluation and Measurement. CEDIL Methods Working Paper 9. London/ Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). https://doi.org/10.51744/CMWP9

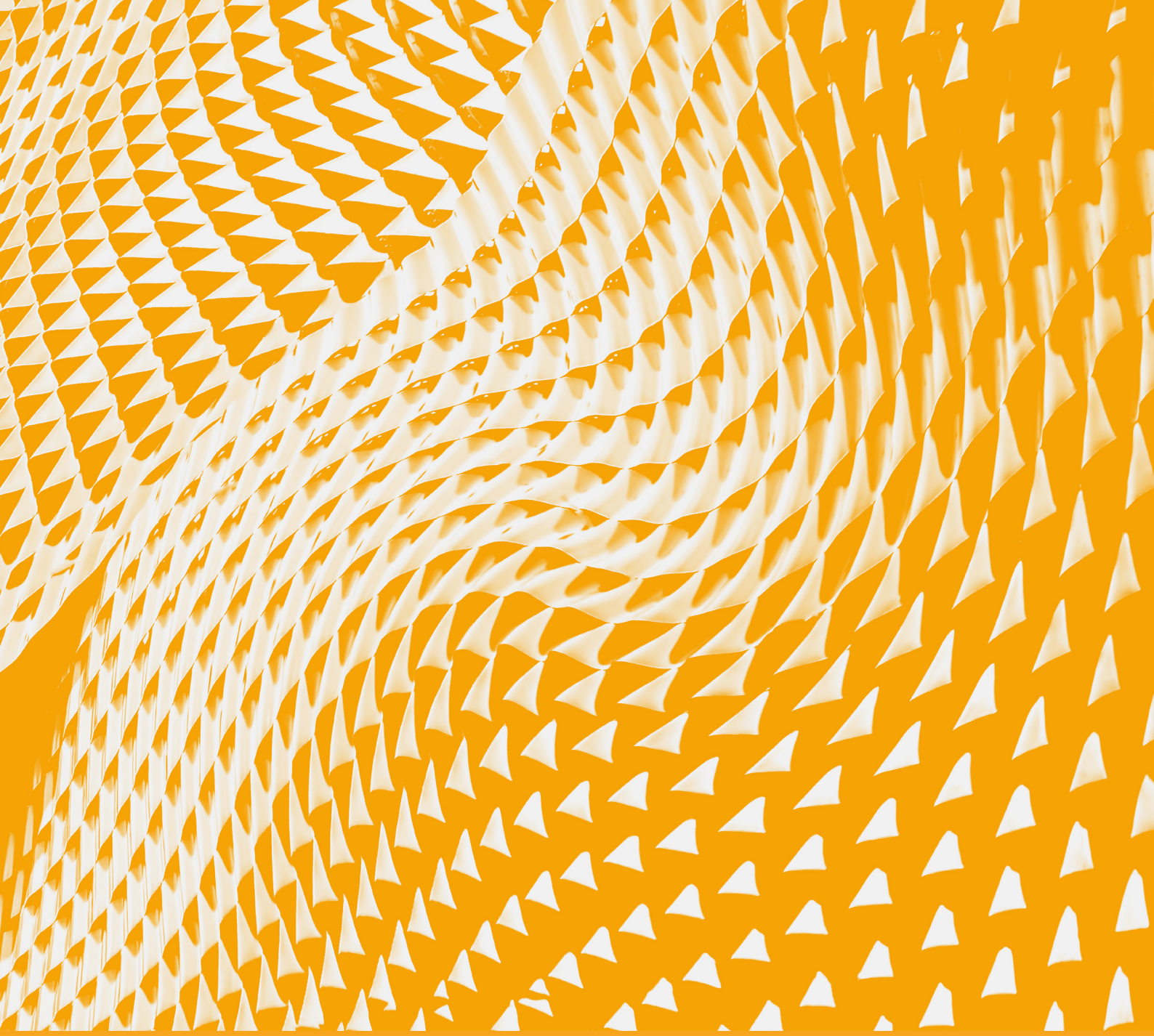
The references listed under Endnotes are recommended for further reading.

Suggested citation

Cavatorta, E. and Attanasio, O. (2022) 'Evaluation and measurement', *CEDIL Methods Brief 9*. London and Oxford: Centre of Excellence for Development Impact and Learning. https://doi.org/10.51744/CMB9

Photo credits

areta ekarafi (p. 2); Inés Fernández (p. 5); CPT Palestine (p. 6); catd_mitchell (p. 8); Kashfi Halford (p. 9); nevil zaveri (p.10); all shared on flickr with Creative Commons licences



@CEDILProgramme

Centre of Excellence for Development Impact

Centre of Excellence for Development Impact and Learning

UKaid from the British people

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK Government through UKaid. The mission of the centre is to test innovative methodologies in evaluation and evidence synthesis and to promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

For more information on CEDIL, contact us at **cedil@opml.co.uk**
or visit our website **www.cedilprogramme.org**