BUILDING EVIDENCE IN EDUCATION

ASSESSING THE STRENGTH OF EVIDENCE IN THE EDUCATION SECTOR

Strong evidence is of central importance in informing policy and programming decisions across all agencies and organisations working with education systems in developing countries. Robust research and evaluation generates the evidence required to form judgements, deliberate options and make intelligent decisions about how to spend scarce financial resources. Programmes taken to scale should be based on rigorous evidence. Assessing and agreeing on the strength of evidence is a challenging task. It requires a combination of technical knowledge and analytical skills to judge research against agreed criteria. This guide provides an introduction to the appraisal of the quality of individual studies and the assessment of the strength of bodies of evidence.

The BE^2 Steering Committee*

Cover Design: JBS International, Inc.

Cover Photo: DFID

* The Building Evidence in Education (BE^2) working group is led by a Steering Committee composed of the Department for International Development (DFID), United Nations International Children's Emergency Fund (UNICEF), United States Agency for International Development (USAID) and The World Bank Group.

Assessing the Strength of Evidence in the Education Sector

Foreword

The Building Evidence in Education (BE2) donor working group was launched in 2012 with the aim to engage bilateral and multilateral donors and foundations committed to:

- increasing the quality of education research;
- promoting the use of evidence in education programming; and
- strengthening donor research collaboration.

The working group is led by a Steering Committee composed of the Department for International Development (DFID), United States Agency for International Development (USAID), The World Bank Group and a rotating representative of the United Nations (UN) organizations, currently the United Nations International Children's Emergency Fund (UNICEF).

This series of Guidance Notes,, prepared for the BE2 working group by its respective members, provides tools and guidance for generating better evidence and leveraging existing evidence more effectively and efficiently. These Guidance Notes have benefited from the advice of BE2 member organizations and are valuable tools for researchers and commissioners of research.

Chris Whitty
Chief Scientific Officer
DFID

Josephine Bourne
Associate Director Education
UNICEF

Christie Vilsack
Senior Advisor for International Education USAID

Claudia Costin
Senior Director, Education
The World Bank Group

This guidance note is based on DFID's How to Note on 'Assessing the Strength of Evidence' produced by Will Evans and Mark Robinson. It has been adapted to the education sector through a consultative process by Rachel Hinton, Senior Education Advisor, Research & Evidence Division, DFID, Olivia Faulkner, Monazza Aslam and BE^2 Members, on behalf of the Building Evidence in Education (BE^2) working group. BE^2 thanks all its members for input provided to earlier drafts of this guidance note.

This page intentionally left blank.

Contents

This page intentionally left blank.

A. Introduction

a. Why Strong Evidence Matters

Strong evidence is of central importance in informing policy and programming decisions across all agencies and organisations working with education systems in developing countries. Robust research and evaluation generates the evidence required to form judgements, deliberate options and make intelligent decisions about how to spend scarce financial resources. It is therefore vital that research evidence is evaluated in a fair and balanced way not only for effective policymaking but also to guide further research and to draw informed conclusions.

b. The Purpose of This Guide

In recent years there has been an increased emphasis on the generation and use of rigorous evidence to inform programme design and policymaking. The objective of this is to base decisions on 'what we know' rather than on conjecture and, in doing so, achieve the best value for each pound/dollar spent and enhance the success and impact of policies and programmes. Increasingly, donor agencies and other organisations are commissioning quality research that rests on strong principles of research methodology using different types of data. Significant effort has also gone into identifying 'quality' research evidence – aiming to test assumptions, answer specific questions or test out hypotheses – which can then be relied upon to be rigorous and substantive whilst also being objective and easily available.

This guide provides staff from donor agencies (and researchers and practioners who may be interested) with a thorough introduction to:

(a) the appraisal of the **quality of individual studies** and
(b) an assessment of the strength of **bodies of evidence** in education.

Specifically, the guide aims to help staff and other individuals to understand different types of research and assess its quality in order to determine what can and cannot be concluded from it. More generally, it aims to set out common standards for the international community on how to assess evidence. Agreement on the strength of evidence on particular issues in the education sector is critical in enabling us to speak with a unified voice when we provide policy advice to our government and national counterparts. Donors are typically interested in identifying the key research gaps and drawing out policy recommendations from a piece of research. While individual research studies may not focus on policy implications, donors are committed to ensuring policy is evidence-based and will thus seek clear recommendations emerging from syntheses of the evidence. DFID, for example, typically expects that these recommendations will be summarised and visually represented through evidence maps and evidence briefs.

c. Importance of Different Types of Research

Most research is ultimately grounded in data. For the purposes of this guide, data sources are described as belonging to one of two categories: quantitative data and qualitative data. This guide includes the variety of research designs and methods used in **social science research**.[1]

Quantitative data are typically data that can be expressed numerically. Methods and designs using data of this type use mathematical techniques to illustrate data or explore causal (cause and effect) relationships.

Qualitative data typically involve categorising and classifying information rather than using numerical values. The methods and designs used in interpreting data of this type rely on collating and analysing the resultant rich information to infer meanings. The researcher usually attempts to understand the *mechanisms* behind impact or cause and effect relationships.

If research is all about the quest for 'answers', then the consumers of research are entitled to expect that those 'answers' are credible and trustworthy. This is especially important in studies which seek to explore cause and effect, or action and reaction. Some types of research (discussed later in the guide) explicitly seek to demonstrate cause and effect relationships and are able to do so with varying degrees of confidence. Some types of research are especially good at reducing the risk of bias and may, therefore, be seen as the 'gold standard' of research aiming to isolate cause and effect. However, other research can bring depth to our understanding of why some events unfold as they do, and critically help us understand people's behaviours, perspectives and interpretations of the events that affect them. This is often where certain research approaches (discussed below) add substantial value.

In recent years, researchers have also recognised the value of **mixed methods** research, which uses more than one method of data collection during a research study and typically involves mixing quantitative and qualitative data collection and interpretation approaches. The rise of behavioural economics in recent years is testament to the value of bringing different disciplines and methodologies together to better understand human behaviour. This guide emphasizes the value of all these approaches without accentuating one over the other.

Assessing the strength of evidence is a challenging task. It requires a combination of technical knowledge and individual judgement. It is also likely to require consultation with subject experts and with colleagues who have specialist knowledge. The new research network in education, Building Evidence in Education (BE^2), is committed to promoting these issues and highlighting best practices in capacity building to help all practitioners to improve in this task over time.

[1] The Economic and Social Research Council includes the following disciplines as social science research: economics, psychology, political science, sociology, anthropology, geography, education, management and business studies, though some subject areas (such as livelihoods) cut across the social and natural sciences.

d. A Note on Terminology

Note that the terms ‘quality’, ‘size’, ‘context’, ‘consistency’ and ‘strength’ of evidence should be used with care in accordance with the definitions in this guide.

B. Describing a Single Study

The guide recommends that single studies be described and categorised by type, design and method. The sections that follow explain how.

a. Type of Research

This guidance note recommends the categorisation of research studies by overarching type as follows:

- *Primary and empirical* (P&E) research studies observe a phenomenon at first hand, collecting and/or analysing 'raw' data.
- *Secondary* (S) research studies review other studies, summarising and interrogating their data and findings.
- *Theoretical or conceptual* (TC) studies, like secondary research studies, draw on previous research, but they do so primarily to construct new theories rather than fresh empirical 'evidence'.

Summary : Research Type

Research Type	Definition
Primary & Empirical (P&E)	Observe a phenomenon at first hand, collecting, analysing or presenting 'raw' data
Secondary (S)	Review other studies, summarising and interrogating their data and findings
Theoretical or Conceptual (TC)	Drawn on previous research also but primarily to construct new theories rather than fresh empirical 'evidence'

b. Research Design

This note also recommends the categorisation of research studies according to research design (see Boxes 1).

Box 1: What Is a Research Design?

A research design is a framework in which a research study is undertaken. It employs one or more research methods to (a) gather data and (b) analyse data. Both conventional research studies, and evaluation studies (such as impact evaluations) employ research designs and methods to gain insights into the real world. The gathering/collection and the analysis of data can involve quantitative or qualitative approaches or both.

Many (but not all) research designs aim in some way to explore causal relationships: 'What is the size of the effect of x on y?' or 'why does x cause y?' However, some research aims only to identify the *association* between one variable and another (see Box 2).

Box 2: Causal and Correlational Relationships

All relationships in analyses are aimed at identifying the association between any two variables. However, while a **correlational relationship** simply indicates that the two variables have *some* association (working either positively or negatively together), a **causal relationship** exists when one variable clearly *causes* another. For example, it is often claimed that higher ability individuals also perform better in mathematics assessments. In general, people who are more able may also have a higher tendency to be good at mathematics. However, knowing that the two variables are correlated does not tell us that higher ability necessarily *causes* better performance in mathematics.

Different designs are more or less appropriate for teasing out alternative aspects of such causal relationships. Different designs are also more or less suited to exploring the wider applicability of the research findings to a variety of contexts.

Different research types adopt varying research designs. Typically, primary and empirical (P&E) research types employ the following research designs:

- Observational/descriptive or non-exprimental (OBS) research designs
- Quasi-experimental (QEX) research designs
- Experimental (EXP) research designs

Secondary research usually involves the following designs:

- Systematic reviews
- Rigorous reviews
- Non-systematic reviews

Full explanations of each of these categories and descriptors is provided in the relevant sections below.

This guide deliberately avoids constructing a hierarchy of research designs and methods. Instead, it recognises that different designs are more or less appropriate to varying contexts and answer differing research questions.[2] For example, an experiment will answer whether or not there is an impact, whereas a non-experimental study may be more useful for unpacking if, why, how and for whom there is an impact. The use of different types of research design may also be determined by the type and quality of data available.

[2] Stern, E., N. Stame, J. Mayne, J. Forss, R. Davies and B. Befani. *Broadening the range of designs and methods for impact evaluations.* DFID Working Paper 38. London: Department for International Development, 2012, p. 2.

Counterfactuals – measuring what would have happened in the absence of an intervention – are important for establishing a causual relationship. A counterfactual can be created in a number of ways, with impact being estimated by comparing outcomes in the absence of the intervention with those under the intervention. Non-experimental designs are important for explaining the nature of, and mechanisms behind, those relationships. Typically, stronger bodies of evidence are likely to be characterised by the availability of a wide spectrum of evidence which uses and triangulates several research designs and methods. In recent years, researchers have also recognised the value of 'sequential research designs' in which observational/descriptive designs, which are typically cheaper to implement, are used before more expensive experimental designs are used to tease out causal relationships.[3]

Research Designs Used in Primary and Empirical (P&E) Research

Observational/descriptive (OBS) designs or non-experimental designs encompass a wide range of valid empirical **methods** (discussed below), designed in different ways to answer different questions. Some designs within this subgroup of empirical research aim to explore causal relationships. They may be concerned with the effect of a treatment (e.g. a drug, a herbicide) on a particular subject sample group, but the researcher does not deliberately manipulate the intervention and does not assign subjects to **treatment** and **control** groups (Box 3). However, an OBS design may collect data in non-treatment areas, as a means of understanding causality.

Box 3: Treatment and Control Groups

A **treatment group** is the group/sample in the analysis on whom a 'treatment' or intervention/programme is administered. The **control group**, on the other hand, is identical to the treatment group in all respects except that it does not receive the treatment/programme.

For example, a policymaker or researcher may be interested in identifying whether provision of school uniforms to girls improves their participation in school. To do so, the researcher may administer the intervention – uniforms – to a subsample of girls. This subsample forms the 'treatment' group. The 'control' group is the subsample of girls to whom uniforms will not have been administered.

Treatment and control groups can form part of observational (OBS), quasi-experimental (QEX) and experimental (EXP) research in different ways. In some types of OBS research, 'treatment' and 'control' groups can be identified by the researcher who does not manipulate the intervention by assigning subjects to treatment and control groups. In QEX techniques, while the intervention may be administered to a 'treatment' group, it is not administered to a randomly selected control group. Instread, a comparison group is created statistically. However, only in EXP designs do researchers *randomly* allocate subjects to treatment and

[3] Other kinds of analysis – such as policy briefs and articles – that do not necessarily adopt empirical research designs or systematically or non-systematically review them should typically be excluded from review. Any valuable insights from these types of analyses, can, however, be included in the discussion, but they should not constitute the main body of evidence being reviewed.

control groups.

In instances where the researcher/policymaker is especially interested in identifying true cause and effect relationships (i.e. causal relationships), research designs that allow comparisons between treatment and control groups may be especially helpful. However, in some instances, researchers may simply be able to observe the effects of a treatment post hoc or may be interested in answering other questions regarding the sample group. In such a case, it is sensible to adopt research designs that reflect the needs of the researcher.

In some instances, designs within this framework may be aiming to analyse patterns and behaviours, without necessarily attempting to demonstrate the size or strength of a causal linkage. Research methods typically used within this class of research design include (and are not limited to): case studies, historical analyses, theory-based analyses, ethnographies, participatory designs, and more quantitative analyses using cross-sectional or panel data (see below).[4]

Quasi-experimental (QEX) Research Designs also involve the observation of control groups as compared to treatment or intervention groups. However, in a quasi-experimental design, an intervention is administered to a treatment group but not to a control group, and the resulting differences between the two groups are observed. In quasi-experiments, the researcher may or may not have discretion over the assignment of the treatment. Even in cases where they do, subjects in quasi-experiments are not allocated to these groups *at random*. This differentiates them from experimental designs and somewhat reduces the confidence with which any effect can be attributed to the intervention. Quasi-experimental studies often use statistical analysis to compensate for the potential biases of non-randomisation and to bolster the construction of a robust counterfactual argument (see above). Case-control studies and regression discontinuity design are examples of quasi-experimental methods.[5]

Experimental (EXP) designs are typically concerned with the effect of a treatment/intervention or programme on a specific group. The treatment is not given to a control group, and the resulting differences between the two groups are observed using inferential statistical analysis. This enables the construction of a robust counterfactual argument (i.e. 'What would have happened in the absence of x?' – i.e. in the absence of the treatment/programme). Crucially, experimental designs allocate subjects (people, animals, villages, etc.) to treatment or intervention groups *at random*. This increases the chances that any difference in effect observed is a direct result of the treatment administered. Such designs are useful for demonstrating the presence and magnitude of causal linkages and hence attribution (e.g. '*a* causes *b*') with a higher degree of confidence. Randomised control trials (RCTs), a type of research method within this class of research design (discussed below), are a well-established form of experimental research. Experimental designs are considered a 'gold standard' for addressing certain types of questions but may not always be considered

[4] Stern, E. et al., 2012.
[5] See, for example, White, H., and D. Phillips, *Addressing attribution of cause and effect in small n impact evaluations: Towards an integrated framework*. 3ie Working Paper 15. New Delhi: International Initiative for Impact Evaluation, June 2012.

appropriate in international development. This is because the methods using these designs typically require randomly assigning a large number of individuals into treatment and control groups, which can make the methods costly, time-consuming and sometimes unethical. Moreover, what appears to work for one group of individuals or in a given region (as shown by an experimental design) may not be generalisable for the entire population, raising questions about 'external validity' (see below).

Research Designs Used in Secondary (S) Research

Systematic review (SR) designs adopt systematic methods for searching for and synthesising literature on a given topic. They interrogate multiple databases and search bibliographies for references. They screen the studies identified for relevance, appraise their quality (on the basis of the research designs and methods they employ), and synthesise the findings using formal quantitative or qualitative methods, or both. Systematic reviews are always clearly labelled as such.[6] They represent a robust, high quality technique for evidence synthesis. Yet some caution should be exercised in interpreting the findings from systematic reviews: the treatments and outcomes that they summarise are not always similar enough across studies to allow for meaningful comparison. Moreover, the synthesis of multiple studies from similar contexts cannot form the basis for generalised claims across other contexts.

Rigorous review (RR) designs typically do not require users to collect as much information about the process as SR designs. They do not require reviewers to keep track of studies that are excluded from the review and are also typically more sensitive to 'information architechure' in that they are more open to inclusion of 'grey literature' and resources outside of peer reviewed journals. Finally, these research designs typically involve greater subjectivity at different steps in the review process, and as such emphasis is placed on ensuring this is documented and acknowledged.[7]

Non-systematic review (NSR) designs also summarise or synthesise literature on a given topic. Some non-systematic reviews will borrow some systematic techniques for searching for and appraising research studies and will generate rigorous findings, but many will not. Policy analyses, evidence papers and rapid reviews may fit into this type of research design.

Theoretical or conceptual (TC) research studies may adopt structured designs and methods but do not generate empirical evidence. Theoretical or conceptual research may be useful in designing policy or programmes and in interrogating underlying assumptions and empirical studies, but should not be referred to as 'evidence'.

[6] http://dfidinsight/Other/Departments/EvidenceResources/Synthesizedevidenceproducts/Systematicreviews/index.htm

[7] Drawn from Hagen-Zanker, J. and R. Mallett, *How to do a rigorous, evidence-focused literature review: a guidance note*. ODI Working Paper. London: Overseas Development Institute, September 2013. Available at http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/8572.pdf

c. Research Methods

Research methods are particular tools or techniques for gathering and/or analysing data in primary and empirical research. A variety of research methods are considered to be valid and capable of producing evidence. DFID's Research and Evidence Division has drawn up a *Guide to Research Design and Methods* (April 2014) which outlines the main families of research method.[8]

To summarise, **data collection** can be either quantitative or qualitative or a mixture of both (mixed methods).

Once data have been collected, the researchers will analyse the data. Typically, **data analysis** methods can be quantitative (use mathematical techniques to explore correlational relationships or establish causal linkages) or qualitative (collating 'rich' data to infer meaning). Increasingly, the line between quantitative and qualitative research is being blurred by the successful development of **mixed method studies**, which analyse qualitative data quantitatively or interrogate quantitative data through a qualitative lens.[9]

Different research methods may be used within any given design – experimental, quasi-experimental or non-experimental – in primary empirical research.[10] The main methods used are summarised below. It should be noted that some of these research methods can typically belong in more than one design. For example, propensity score matching (PSM) methods and double difference (DD) methods (explained below) are often more extended and flexible forms of regression analysis and could, therefore, be classified under observational (OBS) designs. However, some researchers classify them under quasi-experimental designs (QEX). The reviewer should be open to this flexible classification when reviewing studies. Moreover, how data are collected often determines the types of methods that can ultimately be used for analysis. Below, we discuss different research methods for collecting data, followed by some examples of methods that can be used in the analysis. It should be noted that these data collection methods may not be mutually exclusive.

Research Methods Typically Used to Collect Data

Large*-n *surveys usually involve collecting quantitative or qualitative data from a cross-section sample of a population at a single point in time, producing a 'snapshot' of a population or society. Cross-sectional data can be gathered simply for descriptive purposes, outlining the parameters of particular subjects (a population, group of countries, etc). Cross-sectional data can also collected for more intense quantitative analysis, where the effects of one characteristic (variable) of the population or group on another are tested. The sample sizes for these types of analyses can be large surveys (single observations from a sample survey based on random or other sampling techniques or all units in the population, i.e. a

[8] The *Guide to Research Design and Methods* makes clear the distinction between research design and method. It also explains a number of the most commonly used research methods so that reviewers of academic papers can recognise a particular method (and understand its relative merits) when they see it.
[9] Stern, E., et al., 2012, p. 30.
[10] Unlike primary research, secondary research designs are typically reviews or syntheses of others' research.

census). Cross-sectional studies are comparatively low-cost, can examine multiple subjects/variables and, when based on a representative sampleof the population, can be generalised to the whole population. These research methods typically aim at testing specific hypotheses, such as 'is socioeconomic status associated with poorer learing outcomes among school age children?' or 'is going to a private school associated with better learning outcomes as compared to going to a government school?'.

Cohort/longitudinal/panel data collection methods differ from cross-sectional methods in that they collect data on the same unit of observation (individual for instance) for at least two periods of time (sometimes several years, even decades). Sample sizes vary but are usually large. Longitudinal *cohort* studies gather data from the *exact same sample group* over time. These research methods are also typically interested in testing hypotheses and, because of the nature of the data, are able to use more stringent and robust empirical methodologies to arrive at conclusions.

Interviews/focus groups involve research methods which can use a variety of techniques, such as semi-structured or structured interviews administered to individuals or focus groups. These can involve group interviews with a number of people on a specific topic or issue.The sample sizes used in these methods are typically small (focus groups usually have at least four or five respondents), but these methods generate rich qualitative information on the opinions, attitudes and feelings of the subjects being surveyed and are helpful in answering the 'why' and 'how' types of questions that may interest researchers.

Ethnographic research involves the study of social interactions, behaviours and perceptions that occur within groups, teams, organisations and communities. The resulting sample size is usually small, often even a single case. This approach typically involves the investigation of a few cases, sometimes even just one case, but in detail. The emphasis of this type of research is to explore social phenonema rather than test specific hypotheses.

The ***case study method*** of data collection involves a descriptive, exploratory or explanatory analysis of a person, group or event. In-depth analysis of the subject – i.e. the 'case' – provides the analytical framework within which the study is conducted. This method usually results in a non-representative and small sample but is likely to yield rich qualitative data, again aimed at exploring social phenomena rather than testing specific hypotheses.

Randomised control trials are the most common research method used to collect data within experimental research designs. This method involves two key features by design:

- *Manipulation of an independent variable* (provision of an education intervention such as uniforms or textbooks or the reduction in class size through the provision of an additional teacheretc.)
- *Random assignment* when two or more groups are randomly selected from the same population. This is the key feature of randomised evaluation. Participation in the programme is random.

Sample sizes are usually not very large due to the costs involved in implementing these techniques (though individual observations may run in the hundreds if not a few thousand).

Whilst smaller-scale research (often using qualitative approaches such as case study design and ethnographic research) can be undertaken in an experimental setting, it is not generally considered to be experimental evaluation.

Non-systematic reviews used in secondary analysis are summarized in Box 4.

Box 4: Research Methods Used in Secondary Analysis: Non-systematic Reviews

Within the broad research design titled 'non-systematic review', researchers may come across **evidence papers** (such as DFID-produced synthesis products that may be peer reviewed and borrow a systematic approach to searching and assessing the quality of evidence on a given topic), **literature reviews** (collecting and synthesising literature on a given topic), **rapid reviews** (providing a quick review of easily accessible evidence on a topic) and **policy analyses** (that review literature on policies).

Table A1 in the Appendix summarises some of the key design features of the different research methods in primary and empirical research, highlighting when each is most appropriate to use and including some examples from the education sector where these different approaches have been used.

In the education sector, there has been a rapid rise of high quality quantitative studies. Attention should be paid to the possibility of publication bias, a tendency to report results that are positive differently from those that are negative (i.e. supporting the null hypothesis) or inconclusive. Given the challenge of generalisability and cultural sensitivity, replication studies are important to ensure that the findings in one context are applicable in another.

Increasingly it is also recognised that **sequential data collection methods** – which involve undertaking inexpensive qualitative research first, followed by more expensive evaluations such as RCTs – should be used to identify the key issues before expensive methods are implemented, especially to avoid bias in assessing cultural and other aspects. Qualitative information (such as the socio-cultural context) is essential to gathering sound and valuable quantitative data. For example, when doing an evaluation of a programme, the collection of qualitative data may help identify the key mechanisms through which the programme may be having an impact. Therefore, to gain a more comprehensive picture, sequential methods may be the best way forward. Recent work being done by researchers at the Abdul Latif Jameel Poverty Action Lab (JPAL) showcases the usefulness of these approaches.[11]

[11] Rebecca Thornton and Emily Oster's work in Nepal (http://www.nber.org/papers/w14853) that investigated the link between menstruation and education started off by conducting formative research to gather insights. Using in-depth interviews, the researchers investigated the use of menstrual cups among nurses and other women and the feedback from this was used to design a pilot quantitative study in four schools with 200 girls with a total budget of less than 75,000 USD. The study found a very small association between menstruation and school attendance in this setting, and no significant effect of providing sanitary products on reducing the effects of menstruation. Once the results of this 'small-scale' pilot were analysed and presented, the researchers

Observational Designs: Research Methods Used to Analyse Data

Data collected in large-*n* surveys or over a period of time (in the form of panel data), or through interviews/focus groups, ethnographic or case study method will rely on observational designs for analysis. The ultimate method used for **analysis** will depend on whether the data that have been collected are of a quantitative or qualitative nature.

Regression Analysis: Typically, quantitative data can be analysed using regression analysis (such as ordinary least squares, OLS regressions) or more sophistated techniques such as instrumental variable (IV) approaches, 'fixed-effects' estimation and 'Heckman correction' approaches among others.[12]

Political economy analysis typically involves country case studies and specialist data collection requiring macro-level country analysis, sector-level analysis and problem-driven analysis to gain a deeper understanding of a specific country's political context.

Mixed methods research, as mentioned above, are gaining popularity. They involve the use of embedded design of both qualitative and quantitative techniques of different types. The design might be concurrent (both quantitative and qualitative elements simultaneously) or sequential (one design before another). Sample sizes of quantitative data may be large. Mixed-methods research aims both to test hypotheses and to explore phenomena and understand issues in more detail.

Quasi-experimental Designs: Research Methods to Analyse Data

Sometimes large-*n* surveys or panel data of a quantitative nature allow for the use of more sophisticated statistical techniques for analysis. Some examples are detailed below:

Propensity score matching (PSM) involves constructing a statistical comparison/control group based on the probability of participating in a 'treatment' on the basis of 'observed characteristics'. Participants are matched to non-participants on the basis of this probability (or propensity score). The idea is to artificially create something akin to 'treatment' and 'control' groups by mimicking randomisation and finding from a large group of non-participants those who are observationally similar to participants. Large cross-sectional sample surveys can be used for this type of analysis.

Double difference (DD) methods are unlike PSM methods that are premised on participation being dependent on observed characteristics: DD methods presume that unobserved factors also determine whether an individual participates in a programme or not. However, if these unobserved factors do not vary over time, empirical techniques can be used to 'difference'

---

were approached by several funders to continue this line of research. One of the reasons that the researchers decided not to do so was that they hypothesized similarly small effects of sanitary product provision relative to other possible inputs.

[12] For an extensive discussion of different methods and approaches used within quantitative analysis, see Khandker, S.R., G.B. Koolwal and H.A. Samad, *Handbook on impact evaluation: quantitative methods and practices*. Washington, DC: World Bank, 2010.

their bias out. These methods usually require comparison of participants before and after intervention, so require baseline and follow-up surveys of large-scale data (i.e. data collected at two points in time for same individuals). Repeated cross sections can also be used to do DD.

Regression discontinuity designs exploit some natural variation or delay in programme implementation (based in eligibility criteria or other exogenous factors). They therefore rely on identifying cases 'just above' and 'just below' a given threshold based on the notion that they are likely to be 'similar'. For example, to measure the impact of Grameen Bank loans targeted to households with landholdings of a certain value range, identifying a sample of households just above and below this landholding size presumes that impact of the programme can be identified as households just above and below the threshold are similar to those targeted. Samples can be based on any survey (cross-section or panel) with identification of sub-samples that have been 'exposed' to programme and 'not exposed' to the programme. These designs may also involve comparing cases where interventions were in a 'phased-out' nature.

Experimental Designs: Research Methods to Analyse Data

Both quantitative and qualitative data may be collected within an experimental setting, and different methods for analysis can therefore be adopted depending on the type of data collected. Typically, quantitative data can be analysed using fixed effect and DD methods (but this time using data on the treatment and control groups). In this sense, it has been noted that the data generated using an experiemental design is ultimately analysed using the same statistical techniques often used in non-experimental or quasi-experimental designs, which make the analysis open to similar issues and biases. Qualitative data generated within an experimental setting can be analysed accordingly but, as before, is only likely to allow rich inferences rather than the establishment of causal relationships.

Using Descriptors to Describe Individual Studies

This note recommends the use of the following descriptors to describe single research studies by type:

Research Type, Design & Method

Research Type	Research Design and Method
Primary and empirical (P&E)	Typical methods that may be used in observational (OBS) designs: • Cross-sectional regression analysis/large-*n* survey regression analysis • Cohort/longitudinal/panel data regression analysis • Analysis of interviews/focus group data • Analysis of ethnographic research • Case study research analysis • Political economy analysis • Mixed-methods research

Research Type	Research Design and Method
	Typical methods that may be used in quasi-experimental (QEX) designs: • Propensity score matching • Difference in difference • Regression discontinuity design
	Typical method that may be used in experimental (EXP) designs • Difference in difference
Secondary (S)	Systematic review (SR)
	Rigorous reviews (RR)
	Non-systematic review (NSR) designs: • Evidence papers • Literature reviews • Rapid reviews • Policy analyses
Theoretical or conceptual (TC)	N/A

This note recommends that the researcher clearly indicate the research type, design and method on which a single study is based. In practice, synthesising evidence using this convention would result in summaries of single studies illustrated by the following examples:

- When citing a primary and empirical study by Jones, who uses a quasi-experimental research design with propensity score matching methods, the citation may be written as (Jones, 2005 [P&E; QEX, propensity score matching]).

- In the case of a non-experimental case study by Smith, the citation may be written as (Smith, 2004 [P&E; OBS, case study]).

- In the case of a secondary study by Vaughan, where it is clear that a formal systematic review design was employed, the citation may be written as (Vaughan, 2008 [SR]).

C. Assessing the Quality of Single Studies

Following the description of a single study by type, design and research method used, the reviewer or user should aim to consider its quality. Although this is not a trivial exercise, there are some general rules of thumb that can assist in this exercise. It is also important to note that much of the discussion below allows researchers to assess the quality of primary and empirical research. The quality assessment of secondary studies is somewhat different as is discussed briefly in Box 5 below.

Box 5: Assessing Quality of Secondary Studies

In assessing the quality of secondary studies, a researcher should aim to ask the following questions:

- Does the author state where they have searched for relevant studies to be included in the review?
- Does the author attempt any quality assessment of studies they found?
- Are the study's findings demonstrably based on the studies it reviewed?

Because they address all of these issues directly, peer reviewed systematic reviews can be assumed to be of a high quality. Rigorous reviews may not be as stringent as systematic reviews but are often of a moderate-high quality.

For further guidance about what high quality secondary reviews look like, see Hagen-Zanker, J. and R. Mallett, *How to do a rigorous, evidence-focused literature review in international development.*[13]

Source: DFID Research and Evidence Division, *Research methods guide, evidence into action team*. London: Department for International Development, April 2013.

The reviewer is looking principally to assess the quality of the study *in its own right* and its appropriateness for answering the research question posed by the author of the study. An assessment of the *relevance* or applicability of the study to a specific policy question or intervention design is an important but separate part of evidence synthesis and is covered later in this guide.

a. Proxies for Quality: Journal Rankings and Citation Frequency

Rankings and rating systems applying to both journals and individual academics can provide a useful proxy guide to the quality of a research study, although the validity of such rankings for such purposes is subject to considerable debate. Journal rankings provide an indication of the standard of peer review to which a publication has been subjected, or information on the

[13] Available at http://www.odi.org.uk/publications/7834-rigorous-evidence-focused-literature-review-international-development-guidance-note

frequency with which a study or academic has been cited.[14] The status of publications, in terms of the impact factor of peer reviewed journals, can therefore inform an assessment of quality. Academic peer-review should therefore be treated as an important mechanism. However, *not all well-designed and robustly applied research is to be found in peer reviewed journals, and not all studies in peer-reviewed journals are of high quality*. Journal rankings do not always include publications by southern academic organisations or from online journals, so a broad and inclusive approach is required to capture all relevant studies.

b. Principles of High Quality Studies

Whilst this guidance acknowledges the diversity of methodological approaches of multiple academic disciplines, it outlines principles of credible research enquiry that are common to all. It also recognises that an assessment of the quality of a social science study should involve consideration of the relationship between the researcher and the subjects being studied and assurance that appropriate ethical guidelines have been followed.[15]

The following principles are ***required*** features of high quality studies. They may be covered explicitly or implicitly by the author of a single study. They include **conceptual framing**, **openness and transparency**, **robustness of methodology**, **cultural appropriateness/sensitivity, validity, reliability and cogency**.

c. Conceptual Framing

High quality studies acknowledge existing research or theory and make clear how the current/new analysis sits within the context of existing work. They typically construct a conceptual or theoretical framework, which shows how a researcher thinks about an issue and lay bare the researcher's major assumptions. High quality studies pose specific research questions or hypotheses to which the research seeks to respond. When researchers and evaluators have embedded a clear theory of change in their work, it makes it easier for researchers to look across sets of studies and draw out knowledge in a meaningful way.[16]

d. Openness and Transparency

What it means: High quality studies should be transparent about the design and methods that have been employed as well as the data (and resultant sample) that have been gathered and analysed. This allows for the study to be repeated and corroborated. Failure to disclose the data and code on which analysis is based raises major questions about the credibility of the

[14] See, for example, the Thomson Reuters' impact factor ratings for 'planning and development studies': http://thomsonreuters.com/products_services/science/academic; Thomson Reuters, *Essential science indicators*; and Thomson Reuters' *Highly cited researchers* index: http://www.highlycited.com/

[15] See, for example, International Institute for Environment and Development (IIED), *Towards excellence: policy and action research for sustainable development*, London: IIED, 2012.

[16] Rachel Glennester's work at the Abdul Lateef Jamil Poverty Action Lab in particular provides an example of the kind of analysis on evaluations of learning interventions in international development that incorporates clear theories of change in the research to draw out meaningful inferences.

research, as other researchers need to be able to reproduce the results easily and experiment with alternative formulations.

How to assess openness and transparency: An important sign of quality is whether the author is being self-critical – being open about limitations and alternative interpretations and pointing out inconsistencies with other results. There is also the question of independence: a study paid for and/or conducted by an aid agency might be perceived as less independent than a study conducted by a third party. A study that explicitly states who has funded/commissioned the study would be considered more transparent than one that has not. The study should also clearly state the sample size and any limitations therein.

Assessing the Strength of Evidence of Secondary Research:
A Non-systematic Review

Langthaler, Margarita (2013). Argumentation Framework: The Effects of Education on Development. Eschborn/Bonn: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH [NSR]

This literature review analyses the current academic debate regarding the effects of education on areas and sectors relevant to development policy and derives key findings for development cooperation. The resulting 'argumentation framework' is designed to help policymakers weigh up the pros and cons of investing in education support as well as its form. A total of 43 publications were considered for this purpose, of which there were 16 quantitative, 5 qualitative and 3 mixed quantitative–qualitative empirical studies. A further 16 literature reviews and 3 synthesis reports were also included in the analysis. However, considering the breadth of the aspects to be studied, ranging from the effect of education on health, economic growth and income to democracy, gender or nutrition, it is realistic to categorise the body of evidence to be of medium size.

Though the study describes the specific fields and sectors that were studied to **identify literature**, it does not clearly indicate the databases that were searched. The **quality** of the literature was evaluated based on academic stringency, professional and/or political influence of the respective publications and their representativeness of a broad spectrum of conceptual and methodological approaches; it can therefore be considered as high. At the same time, methodological and conceptual restrictions of the selected literature are pointed out. Firstly, many of the quantitative studies, in particular those relating to the economics of education, have theoretical foundations which are restricted in disciplinary terms. The variables used, therefore, sometimes raise doubt as to whether they actually reflect the phenomenon to be studied. Secondly, doubts also arise as to how to deal with causal mechanisms which in many quantitative studies relate only to the correlations that are measured without any thorough examination of the respective context. Finally, the comparability of findings across sectors is subject to a terminological restriction. Studies looking at the effects of education from an

economic perspective inevitably employ different educational terms and concepts (education as an economic input variable) to studies which, for example, analyse the relationships between female literacy and empowerment (education as a tool for empowerment). The study is **transparent** in identifying the key methodological weaknesses and the **findings of the study are demonstrably based on the studies that have been reviewed**.

e. Robustness of Methodology

What it means: This refers to the appropriateness of the design and methods to the research question and its rigorous application. There are three main types of design (observational, quasi-experimental and experimental) and many types of methods (discussed above and see Table A1). None is necessarily 'better' or 'worse', but some designs and methods are certainly more appropriate for use in specific settings or for responding to particular types of research questions than others.

Typically, experimental and quasi-experimental research designs tend to be more appropriate for identifying, with confidence, the presence of causal linkages between observable phenomena. Only experimental designs enable clear attribution; non-experimental designs will only enable contribution, but, depending on the robustness and rigour of the design chosen, an assessment of attribution "to the greatest extent possible" will be possible (White and Phillips 2012). However, if the methods are improperly applied, it is possible for experimental or quasi-experimental studies to be of a low quality. For example, the quality may be compromised if inappropriate proxies are used to measure the presence or absence of a phenomenon, or if an experiment fails to take account of political, social or gender phenomena which can bias the findings. The diverse array of non-experimental designs may be more appropriate for contexts and phenomena which cannot easily be explored through experimental and quasi-experimental designs, such as exploring the mechanisms behind a causal linkage, or for deepening understanding of people and behaviours that lie at the heart of most development processes.[17] Crucially, using an inappropriate method to generate data in a particular context is unlikely to yield credible or useful results.

How to assess appropriateness and rigour: The reader of the single study should try to identify the specific question that the paper's author is trying to address. Is it about identifying causation? Is it about quantification of a trend, or about the meaning and implications of a trend? Is the research based on developing a conceptual model and then confronting that model with the data? Answering such queries is a good starting point in determining whether or not the research design and methods employed were appropriate to the study question and context.

[17] See in particular Stern, E. et al. (2012). *Broadening the range of designs and methods for impact evaluations*. Department for International Development, Working Paper 38, pp. 18, 24.

Assessing Appropriateness and Rigour: An Experimental Case Study

Blimpo, M.P., N. Lahire and D.K. Evans (2014). School-based management and educational outcomes: lessons from a randomised field experiment, Unpublished manuscript. Washington, DC: World Bank [P&E; EXP]

This article summarises the results of a gender impact evaluation study conducted between 2008 and 2011 in The Gambia. The study observed the impact of comprehensive school-based management and the capacity building program – Whole School Development (WSD) – on schools, principals, teacher representatives and communities. The research tested receiving a grant only compared to receiving a grant coupled with the WSD intervention that included training in school leadership, management, community participation, curriculum management, teacher professional development, and teaching and learning resources on the school, student, and teacher level. Both were compared with a control group of schools that received no grant nor training. The WSD had no impact on learning outcomes in maths or English. However, the intervention led to a significant reduction in student absenteeism by 21 percent and teacher absenteeism by 23 percent. There was no effect of the grant-only intervention on test scores or participation. The WSD programme does appear to have had positive impacts on test scores in areas with higher literacy levels at baseline. The impact of the programme on enrollment was similar between genders.

Assessing rigour and appropriateness requires answering the following questions: What type of question is the study asking: is it attempting to establish causal linkages? Is the research based on the development of a conceptual model and then confronting the model with the data? Answering these questions will determine whether the research design and methods used were appropriate in answering the question, which can help determine the rigour and appropriateness of design.

The study aims to identify and examine specific *effects* of receiving grants alone compared to receiving grants as well as training on student learning outcomes. The study clearly aims to establish a causal linkage between grants versus grants/training on student outcomes. The experimental design was, therefore, most **appropriate** to answer the research question.

The study demonstrates **rigorous application** of the experimental technique within The Gambian setting. The authors clearly describe the interventions and adopt all the rigours of a well-applied randomisation.

f. Cultural Appropriateness/Sensitivity[18]

What it means: This refers to the appropriateness of the measures/tools/ instruments used by the researchers and their analysis to the cultural context within which the study is set. The extent to which a study takes into account local context has considerable bearing on the way the study is designed, the analytical strategy, interpretation of findings and, ultimately, the quality of the study. This dimension should be assessed separately as it may be that a study is methodologically robust but the measures used or materials and instruments used in a given context are not sensitive to local culture.

How to assess appropriateness/cultural sensitivity: Cultural sensitivity can broadly be defined along two dimensions:

(1) Are the tools/instruments used to measure the impact of the programme culturally relevant?

For all research designs, it is important to consider the extent to which the measures/instruments/variables used in the study suit local contexts. The reviewer should note whether measures have been developed to suit the local context: does the study, for instance, merely translate into a local language or recognise that a test developed in a specific linguistic area may not be automatically suitable to a local context with translation or because of multiple socio-linguistic processes? The reviewer should also note whether local knowledge has been used effectively in the adaptation of measures to reflect resources relevant to the context – for example, are the instruments designed with support and recognition from the local community? – and whether data has been collected in ways that will not generate bias.

(2) Is the analysis of the study culturally sensitive?

This includes the extent to which the analysis includes *locally relevant* social stratifiers (for example, socio-economic status, gender, rural–urban differences, etc.) and influences which may affect interpretation of results. This could, for instance, include a review of the extent to which cross-cultural and cross-linguistic comparisons have been made a part of the analysis, the extent to which local influences have been used in the interpretation of results or the extent to which the study includes the linguistic context. For example, this could include the extent to which a study on learning trajectories allows for dialect effect or the transfer of skills from certain home languages and not others and so on. Collecting data on voter preferences, for example, needs to be done in isolation and away from local social pressures. Taking social norms and pressures into consideration is also important.

[18] This dimension has emerged from the work undertaken as part of the DFID-funded rigorous review: Nag, S., S. Chiat, C. Torgerson and M.J. Snowling, Literacy, foundation learning and assessment in developing countries: Final report. *Education Rigorous Literature Review*. London: Department for International Development, 2014. Special thanks to Sonali Nag for detailed and helpful contributions in developing this measure for assessing quality of evidence.

Assessing Appropriateness and Cultural Sensitivity: Experimental Case Study

Opel, A., S.S. Ameer and F.E. Aboud (2009). The effect of preschool dialogic reading on vocabulary among rural Bangladeshi children. *International Journal of Educational Research*, 48. [P&E; EXP]

This paper evaluates a whole-class dialogic reading intervention in Bangladesh. The main purpose of the study was to examine the efficacy of a four-week reading intervention among rural pre-schoolers with the aim of increasing their 'expressive vocabulary'. Eighty preschoolers from five preschools were randomly selected to participate in a four-week programme. The treatment group were tested on 170 challenging words before and after the intervention with a view to identify their expressive vocabulary and compared to control group children who participated in the regular language programme. Both groups were read eight children's storybooks with illustrations in Bangla, but the dialogic reading teacher was given a set of 'wh' and definitional questions to enhance children's verbal participation. The mean vocabulary scores of dialogic programme children increased from 26 percent to 54 percent while that of the control children remained at the same level.

The **instruments/tools** used are appropriate and culturally sensitive as it was recognised that all children spoke Bangla at home and the materials and activities of the preschool and of the intervention were designed to be in Bangla. Systematic processes were used to check cultural relevance of measurement items (for example, 'In the absence of any list of age-specific words for these Bangla-speaking children, a list was created of words that fit two criteria: they should be known to grade 1 or 2 children but unknown to preschoolers, and they should be used in the storybooks. A list of difficult words was first selected from the grades 1 and 2 readers. The expressive vocabulary of a sample of first and second graders was tested with these words. Words known by more than 50 percent of them were retained. This new list was then administered to children of several non-participating preschools; words known to most of them were excluded.' p.14). The selection of indicators largely demonstrates that the study is moderately–highly culturally appropriate/sensitive.

The **analysis** is broadly culturally sensitive as it notes, for example, that preschool children may start at low vocabularies because they lack responsive and sophisticated dialogue with adults. The low literacy background in the local context also resulted in preschool children not being sufficiently challenged. The study also noted the perception of local teachers about telling stories to children aged three to five, and this improves the cultural sensitivity of the analysis.

Assessing Appropriateness and Cultural Sensitivity: Non-experimental Case Study

Dang, H., L. Sarr and M.N. Asadullah (2011). *School Access, Resources, and Learning Outcomes: Evidence from a Non-formal School Program in Bangladesh*. IZA Discussion Papers 5659, Institute for the Study of Labor (IZA) [P&E; NEX quantitative panel data]

This study uses stringent empirical techniques and panel data to evaluate educational outcomes (enrolment and test scores) of children in rural Bangladesh. In particular, it studies non-formal schooling models that have been brought into the government folds – the Reaching Out-Of-School Children (ROSC) program inspired by the success of the Bangladesh Rural Advancement Committee (BRAC) model. Operative since 2005, the ROSC model has mainstreamed non-formal education into formal schooling through public finance and at the time of writing this paper had reached more than 15,000 schools and served more than 500,000 educationally disadvantaged children in the poorest Upazilas (sub-districts) in rural Bangladesh. The authors compare educational outcomes of children in ROSC and non-ROSC schools and find that ROSC schools increased enrolment probability by between 9 and 18 percent for children in the age cohorts 6-8 and 6-10. Using standardised tests administered to grade 2 students, the study also finds that children in ROSC schools do as well as those in non-ROSC schools. The study also concludes that ROSC schools are more 'cost effective' because of their ability to operate more efficiently than government schools in many respects.

The **measures/instruments** used are not discussed in sufficient detail, and it is not clear from the study text itself whether they are culturally sensitive or not. (Although the authors have subsequently confirmed that indeed these instruments were culturally sensitive, this information is not available in the article itself.) One of the main measures used – test scores – is not discussed in sufficient detail in the paper, and it is not clear how students were assessed, what the content was and to what extent the assessments were adapted to the local and regional context. It is, therefore, not possible to assess the cultural sensitivity of the measures and instruments appropriately.

The **analysis** is significantly culturally sensitive as it discusses the factors that undermine or promote educational outcomes within the Bangladeshi context. The study discusses the use of two supply-and-demand side interventions – a school-only grant and a school grant plus an education allowance – which the authors discuss in relevance to the context where grants are used to provide key inputs to schools while the education allowance provides a conditional monetary incentive for out-of-school children to attend school. These locally relevant stratifiers are important components of the program and have been well discussed in the analysis.

The selection of indicators largely demonstrates that the study is **moderately culturally appropriate/sensitive**.

g. Validity

What it means: There are several types of scientific validity. Four of the most important are covered here.

- *Measurement validity*: During the data collection phase of a study, a researcher who sets out to measure or interrogate a particular concept typically selects a particular indicator to do so (e.g. metres as an indicator to measure distance). 'Measurement

validity' describes whether or not the indicator is well suited to measure the concept in question. For example, if a study aims to measure individual welfare, it has to use a valid indicator of 'welfare'. Family income, individual health or individual happiness might be valid indicators, but, in contrast, the value of national exports would be much less satisfactory.

- *Internal validity*: Some studies (typically experimental and quasi-experimental designs) seek to demonstrate that the emergence of one factor is attributable to (i.e. caused by) another. For example, a study may show that rich people tend to live in expensive neighbourhoods. But are they rich because they live in a wealthy neighbourhood, or is the causal relationship working the other way round? Assessing the 'internal validity' of a study means evaluating whether or not the technique that the study uses to explore such causal chains is satisfactory. If the design doesn't take account of 'unseen' (sometimes called 'confounding') factors that might be causing a particular phenomenon, then the study may over- or underestimate the importance of a particular issue as a cause of an observed outcome or behaviour. This may result in a bias in determining which variable is the cause and which is the effect. Threats to internal validity may arise due to confounding, selection bias, experimenter bias, etc. (see below).

- *External validity*: This describes the extent to which the findings of a study are likely to be replicable across multiple contexts: Can they be generalised? Does the study include full information on local conditions that would make it replicable in a different context?

- *Ecological validity:* This dimension of validity relates to the degree to which any research is really able to capture or accurately represent the real world without the research itself impacting upon the subjects it seeks to study. Ecologically valid studies explicitly consider how far the research findings may have been biased by the activity of doing the research itself. Such consideration is sometimes referred to as 'reflexivity' or 'experimenter bias'.[19]

Some types of bias compromise internal validity. They are summarized in Box 6 below.

Box 6: Threats to Internal Validity: Types of Bias

The internal validity of a study can be compromised due the presence of either one or all of the following biases:

Confounding bias occurs when changes in the dependent variable cannot be attributed to the independent variable being studied but rather are attributable to a third variable that is related to the independent variable. For example, in a study investigating the relationship between education and women's earnings in the labour market, failure to control for 'innate' ability may result in confounding bias. This is because earnings may be determined by ability as

[19] See, for example, IIED, Towards excellence: policy and action research for sustainable development, 2012.

well as schooling, and failure to control for it will undermine the causality of schooling's relationship with earnings.

Selection bias refers to the problem that differences in groups exist prior to analysis and failure to control for these differences will result in a bias in the relationship that is ultimately observed. For example, in an analysis of women's earnings in the labour market, by focusing only on women who participate in the labour market the researcher is focusing on a 'select' sample of women, the characteristics of whom may be very different from the characteristics of those who do not participate. The resulting relationships will be potentially biased unless methods are used to overcome this challenge.

Experimenter bias occurs when individuals who are conducting the experiment behave differently among treatment and control groups in ways which may potentially affect the ultimate outcome being studied. It may be possible to reduce this bias using 'double blind' studies in which even the experimenter is not aware whether the participant belongs to the control or treatment group.

How to assess validity: In the case of measurement validity, it is important to repeatedly consider whether or not the indicator chosen fully captures the concept being measured. Are there other dimensions of the central concept that are being ignored? In the more complex case of internal validity, a starting point is to try to think of other possible causal mechanisms that the researcher has not acknowledged. In the case of external validity, the reviewer needs to consider whether the case or context being studied is highly particular or is 'generalisable' to multiple settings.

Assessing Validity: A Longitudinal Case Study

Rolleston, C., Z. James, L. Pasquier-Doumer et al., 2013. *Making Progress: Report of the Young Lives School Survey in Vietnam*. Oxford: Young Lives [P&E; NEX quantitative longitudinal]

A recent school survey in Vietnam, conducted as part of the Young Lives longitudinal study, included 3,284 pupils in grade 5 in five provinces of Vietnam. This study assessed students in the key curricular domains of mathematics and Vietnamese reading comprehension at both the beginning and the end of the 2011–2012 academic year. These assessments were specifically designed to enable value-added analysis of the correlates of learning progress on these domains during grade 5.

The study authors took a number of steps to ensure that the tests used to assess student progress measured what they set out to measure, i.e. to ensure **measurement validity**. Items included in the test had to relate directly to what grade 5 children would be expected to know at the start and end of the school year. To achieve this, the study authors:

- Ensured that test items related directly to competencies which children would be expected to have been taught, via a review of textbooks and the curriculum;

- Framed questions using examples and situations that children from diverse backgrounds would be familiar with;
- Presented questions in a format that would, as much as possible, be similar to that which children encountered at school ;
- Extensively piloted the tests in multiple regions of Vietnam with both teachers and students;
- Undertook detailed statistical analysis to assess the internal consistency of questions in order to refine and adjust the assessment tools.

The study authors provided clear descriptions of the steps they took to enhance the study's validity in other ways. For example, in assessing learning progress of pupils in grade 5, the study controlled for student background and estimated models using 'class fixed effects' so that differences between classes were removed and the results only compared children with their peers. By including initial test scores into the estimation, the study was able to derive a 'causal interpretation' of student outcomes as controlling for this proxies for a large number of unobserved variables (such as ability) and background factors that may generate biases. The study is, therefore**, internally valid**. The study is based on longitudinal data collected from 5 provinces out of 58 in Vietnam, the generalisability of the findings is somewhat questionable (**external validity**), and there is no discussion of whether the findings could have been influenced by the process of research itself (**ecological validity**). In these regards the study is somewhat weak.

h. Reliability

What it means: Reliability usually means one of two things. First, a study is reliable if both the right 'thing' is being measured and it is being measured consistently and accurately. Second, an analytical technique is reliable if the analysis produces consistent results during the processing or use of data, when repeated multiple times.

An unreliable measurement instrument could potentially undermine an entire study. 'Test scores' might be the right thing to measure in a piece of research (when studying outcomes, for instance), but if not measured accurately, the study is flawed. The reliability of an analytical technique boosts the robustness of a study. If a different result were produced every time the same data was processed with the same technique, the study would not be reliable.

How to assess reliability: Consider the instrument or indicator being used to measure the concept. Some indicators (like corruption 'scores' based on 'expert judgement') may be particularly prone to unreliability or bias. When assessing the reliability of analytical techniques, consider how any weaknesses in the technique might bias the findings of a study or mean that different results could be produced.

Assessing Reliability: An Observational Case Study

Sailors, M., J.V. Hoffman and B. Matthee (2007). South African schools that promote literacy learning with students from low-income communities. *Reading Research Quarterly (42)*3 (July–September 2007), pp. 364-387. [P&E; OBS – qualitative case studies]

This interpretive study explored the qualities of six high-performing schools that served low-income South African students. The theoretical framework and methodology derived from research on effective schools conducted, for the most part, in the United States. Data consisted of interviews and classroom observations over the course of two collection phases, focusing on experiences and beliefs held at individual schools. Within a case study framework, the authors used a constant-comparison approach and cross-case analysis to identify five broad themes associated with these high-performing schools. These schools were safe, orderly and positive learning environments and were guided by strong leaders and staffed by excellent teachers who had a shared sense of 'competence, pride and purpose' that included high levels of school and community involvement. Not all was perfect in these schools; they struggled with issues of class size, highly qualified replacement teachers, the future of the graduates of their schools, and writing instruction. In spite of these struggles, these schools demonstrated determination, resiliency and purpose.

Assessing reliability: The study acknowledged existing research and **posed a research question**: 'What are the qualities of high-performing schools serving low-income South African learners?' The study related to the thematic research question by identifing a number of aspects of school leadership perceived to support high performance in some schools serving low-income communities.
Reliability: This study used multiple researchers to undertake school observations and interviews; the researchers checked their own conclusions with each other and then cross-checked them against the wider analytical team to analyse between schools. The team ensured that different types of data were collected – observations, interviews and document analysis – to triangulate findings and take into account the variety of possible contexts. The authors also provide a good example of how to enhance the reliability of qualitative analysis: interviews were videotaped and transcribed; details were given about the analytical technique; this included a high degree of reliability checks; and researchers conducted interim data analysis and repeated return to the original data sources when constructing an analytical framework.

i. Cogency

What it means: A high quality study will provide a clear, logical argumentative thread that runs through the entire paper. This will link the conceptual (theoretical) framework to the data and analysis, and, in turn, to the conclusions. High quality studies will avoid making claims in their conclusions that are not clearly backed up by the data and findings.

How to assess cogency: If the principles of good reporting have been followed, the author of a high quality study should 'signpost' the reader through the various sections of the study. The reviewer should try to consider whether he or she would have written the same conclusion or executive summary for the study based on the analysis and results it presents.

A really rigorous review of the evidence on a given topic should give due consideration to each of these aspects of study quality. It is possible to construct checklists, or scorecards to grade evidence. Even when formal scoring mechanisms are not used, reviewers of single studies are advised to keep a record of their observations on the following aspects of a study to demonstrate the basis of their assessment and so it can be shared with other members of staff.

It is possible to assess both quantitative and qualitative research across these dimensions. However, the extent to which a certain dimension will be relevant for a particular type of research design may vary. For example, a given dimension may be more relevant for studies that focus specifically on interventions. Moreover, these dimensions are more applicable to primary and empirical research types than to secondary research (see Box 5 above on assessing the quality of secondary research).

Box 7: The Importance of the Value for Money (VfM)

This refers to the approach of maximising the impact of each pound/dollar spent to improve poor peoples' lives. Ultimately, it is about the researcher being aware of the money that enters into the chain and the resulting outcomes and impact.[20] In this context, it refers to the extent to which the study has an impact proportionate to its cost. The purpose of including this dimension into assessment is to identify the extent to which the researcher has applied the understanding of the relative cost of different research designs in relation to the possible impact and outcomes.

Whilst it may not be feasible to consider Value for Money as a required dimension for assessing the quality of an individual study, it can be considered a ***desirable*** dimension, especially when assessing certain types of research designs. It can broadly be assessed along two dimensions:

(1) Value for money (VfM) of research design

At all times, the reader of a study should be alert to the potential cost-effectiveness of or the VfM of the research design. Different research designs are associated with different costs (and these may also vary depending on context). A research design may be empirically sound but prohibitively costly in that it does not offer good VfM. This does not mean that the the 'cheapest' alternative is assessed to be of the best quality. It requires assessing (1)

[20] See, for instance, White, P., A. Hodges and M. Greenslade, *Guidance on measuring and maximising value for money in social transfer programmes – second edition*. London: Department for International Development, April 2013. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/204382/Guidance-value-for-money-social-transfers-25Mar2013.pdf

whether the researcher is aware of the costs of his/her design, (2) whether the design appears to produce VfM in the given context and (3) whether the intervention is capable of being sustained and scaled up. Studies where researchers have been mindful of these considerations may, for example, have adopted sequential research designs – i.e. conducted qualitative research initially at a relatively low cost before implementing a costly intervention. Moreover, cost-effectiveness of the research design may be highly dependent on the underlying assumptions that researchers make and it is critical that they are transparent about these assumptions in any research they undertake.

(2) Cultural appropriateness of the intervention/study

This relates to the extent to which the study is appropriate to local realities and historical context. This relates to VfM to the extent that in a study wishing to assess the results of a development intervention, an important dimension has to be how socially and historically appropriate the intervention/study itself is. For example, are the roles being assigned to key players (e.g. tribal tensions, caste tensions, teacher unions) taken into consideration? Is the intervention designed to take into consideration cultural factors or realities, or is it merely an 'adaptation' of a successful intervention in a different context? In other cases, the intervention may be appropriate but the *fidelity* of implementation – i.e. the extent to which the intervention was delivered as it was intended – may be weak. Cultural appropriateness also includes considerations such as the extent to which the cost of national reforms/development programmes being researched is adapted to reflect resources relevant to the economic and cultural context. For example, consider a study analysing tablet provision in a context where it is unlikely to be scaled up and sustained, or a programme for the extension of the education system to early years in countries struggling with financing primary provision. This aspect refers only to studies that adopt a specific research design, intervention-based studies.

Summary: Assessing the Quality of an Individual Study

Principles of Quality	Associated Principles	Definitions (Page Number)	H/M/L (High, Medium, Low)
Conceptual framing	Does the study acknowledge existing research?	15	
	Does the study construct a conceptual framework?		
	Does the study pose an appropriate research question?		
	Does the study outline a hypothesis?		
Openness and transparency	Does the study present the raw data it analyses?	16	
	Does the author recognise limitations/weaknesses in his/her work?		
	Does the researcher acknowledge their		

Principles of Quality	Associated Principles	Definitions (Page Number)	H/M/L (High, Medium, Low)
	own subjectivity in the process of the research?		
	Is the data disaggregated (by age, gender, etc.) or representative of the population?		
Robustness of methodology	Does the study identify a research design?	18	
	Does the study identify a research method?		
	Does the study demonstrate why the chosen design and method are good ways to explore the research question?		
Culturally appropriate tools and analysis	Are the instruments/tools used to measure the impact of the programme culturally relevant?	19	
	Is the analysis culturally sensitive?		
Validity	Has the study demonstrated measurement validity?	22	
	To what extent is the study internally valid?		
	To what extent is the study externally valid?		
	To what extent is the study ecologically valid?		
Reliability	To what extent does the study demonstrate measurement reliability?	25	
	Has the study demonstrated that its selected analytical technique is reliable?		
Cogency	Does the author 'signpost' the reader throughout?	27	
	Are the conclusions clearly based on the study's results?		

The following descriptors should be used when assessing the quality of single research studies. Assignment of a particular 'grade' to a study is a matter of judgement for the reviewer. It should be based on consideration of each of the criteria outlined above to ensure consistency of approach across studies.

Study Quality	Abbreviation	Definition
Very High	↑↑	Demonstrates strong adherence to principles of appropriateness/rigour, validity and reliability; strongly demonstrates principles of conceptual framing, openness/transparency, cogency, cultural appropriateness and value for money.
High	↑	Demonstrates adherence to principles of appropriateness/rigour, validity and reliability; likely to demonstrate principles of conceptual framing, openness/transparency, cogency, cultural appropriateness and value for money.
Moderate	→	Some deficiencies in appropriateness/rigour, validity and/or reliability, or difficulty determining these; may or may not demonstrate principles of conceptual framing, openness/transparency, cogency, cultural appropriateness and value for money.
Low	↓	Major and/or numerous deficiencies in appropriateness/rigour, validity and reliability; may/may not demonstrate principles of conceptual framing, openness/transparency, cogency, cultural appropriateness and value for money.

j. How It Is Used in Practice

Returning to the previous examples, if a user of evidence cites a primary and empirical study by Jones, who uses a quasi-experimental method, but the paper is of only moderate quality, the citation may be written as: (Jones, 2005 [P&E; QEX; →]).

In the case of a high quality observational study by Smith, the citation may be written as: (Smith, 2004 [P&E; OBS; ↑]). In this case, it is important to be explicit about the method (not just the design) that has been employed.

Those citing evidence should not confuse studies which present 'evidence of no effect' (i.e. they actually show that 'x' has no effect on 'y') and those which 'find no evidence for an effect' (which means that there may be an effect of 'x' on 'y', but it hasn't yet been identified).

D. Summarising the Main Characteristics of a Body of Evidence

Once individual studies have been assessed and appraised for quality, the reviewer may wish to assess the **overall strength** of the existing body of evidence. This section is intended to help reviewers form judgements about the strength of evidence when identifying, sifting and assessing studies for use in programme design and policy papers.

Bodies of evidence should be summarised in terms of four characteristics:

1. The (technical) **quality** of the studies constituting the body of evidence;
2. The **size** of the body of evidence;
3. The **context** in which the evidence is set;
4. The **consistency** of the findings produced by studies constituting the body of evidence.

a. Quality of the Studies Constituting the Body of Evidence

The quality of a body of evidence is determined by the quality of the single studies that constitute it (see above). Remember, the technical *quality* of the body of evidence is just one discrete component of the overall credibility or strength of a body of evidence (discussed below). For example, it is possible for a body of evidence to be small in size but high in quality.

A summary of the technical quality of the body of evidence should build directly upon prior assessment of the quality of single research studies conducted individually or as part of a secondary study such as a systematic review. When summarising the quality of a body of evidence, similar language should be deployed as when assessing the quality of single research studies, but without needing to use directional arrows:

Quality of the Body of Evidence	Definition
Very High	A large majority of single studies reviewed have been assessed as being of a high quality, strongly demonstrating adherence to the principles of rigour, validity and reliability.
High	Many of the single studies reviewed have been assessed as being of a high quality, demonstrating adherence to the principles of rigour, validity and reliability.
Moderate	Of the single studies reviewed, approximately equal numbers are of a high, moderate and low quality, as assessed according to the principles of rigour, validity and reliability.
Low	Many or most of the single studies reviewed have been assessed as being of low quality, showing significant deficiencies in adherence to the principles of rigour, validity and reliability.

This guide does not advocate for the construction of 'hierarchies' of evidence, which assumes the superiority of one or other method. However, it does suggest that bodies of evidence constituted by a diverse range of robust designs and methods (typically experimental and observational and both quantitative and qualitative in nature) are likely to be stronger than those that are reliant upon just one design, or just one or two methods.

b. Size of the Body of Evidence

Across academic disciplines, there is no 'magic number' of studies that, when exceeded, denotes that a sufficient or adequate amount of research has been conducted on a particular topic. Nevertheless, empirical findings can be strengthened through repetition and corroboration, in the same contexts and environments or in different ones. As such, the presence of one study in isolation, uncorroborated by other findings, is unlikely to constitute a large body of evidence.

The size of a body of evidence is also likely to depend on the research question, research context and subject area. When considering multiple dimensions of a major topic (take student learning as an example), it is useful to record which aspects of that topic (e.g. measurement, impact on different areas, influencing factors, etc.) have received greater attention in the literature than others. This gives a sense of the *relative* size of the body of evidence in a discrete field.

Given the absence of a 'magic number' of studies to denote 'adequacy', it is for the reviewer to decide which of the following terms best describes the size of body of evidence. When doing so, it is good practice to list the number of studies that have been identified.

Size of body of evidence
Large (+ state number of studies)
Medium (+ state number of studies)
Small (+ state number of studies)

c. Context of the Body of Evidence

The reviewers of a body of research should also make some note of the origins and context of the evidence that they are quoting. This is closely related to the issue of external validity (see above), and it is particularly important given that in many development sciences and programmatic interventions, the findings of research may be context-specific.

When determining the applicability of evidence from one context to another, the reviewer or policymaker must take note of the consistency of the results of research, any significant variations in the range of results, and the number of comparable contexts from which evidence has been generated. For example, it is possible for there to be a 'large' body of evidence demonstrating the positive effect of a particular intervention, all of which is generated in just two or three countries. Likewise it is possible for there to be evidence sourced from *many* countries but *not* in the country of greatest interest to a programme

designer or policymaker. Ideally, there will be a convincing body of evidence on the likely efficacy of an intervention *both* globally *and* in the context of particular interest.

The descriptors of the size of the body of evidence are as follows:

Context
Global
Context-specific

d. Consistency of the Findings of Studies Constituting a Body of Evidence

Such is the complexity of social phenomena that it is possible to have a large body of evidence drawn from multiple contexts but which nevertheless offers inconsistent findings. In short, the evidence points 'both ways'.

Synthesising multiple studies according to their quality is likely (though not certain) to generate findings that are more consistent. Consistency in a body of evidence reduces uncertainty.

The descriptors of the consistency of the body of evidence are as follows:

Consistency	Definition
Consistent	A range of studies point to identical or similar conclusions.
Inconsistent (mixed)	Different studies point to a range of conclusions. In some cases, one study will directly refute or contest the findings of another. In other cases, different designs or methods applied in different contexts may simply have produced results that contrast with those of another study.

e. Recap: Summarising the Main Characteristics of a Body of Evidence

When summarising or synthesising evidence, a reviewer should seek to make a comment on the quality, size, context and consistency of a body of evidence but may not be able to assess large numbers of individual studies. Instead, the reviewer might use the following types of conventions:

a. 'There is a large (+ indicate number of studies) body of global, high quality evidence relating to the efficacy of direct budget support in poverty reduction. The evidence consistently suggests significant positive effects.' **Or**:

b. 'There is a medium-sized (+ number of studies) body of moderate quality evidence relating to the poverty reduction effects of empowerment and accountability

initiatives. The evidence relates directly to country X. However, the findings of the evidence are inconsistent (mixed).' **Or**:

c. 'There is a small-sized (+ number of studies) and consistent body of evidence that suggests the spread of information and communications technologies (ICTs) is generating greater pressure for increased transparency in government. However, the evidence is of generally low quality.'

An example of how this technique has been used to summarise or synthesise evidence is provided by the table below. This table summarises the use of these categories in arriving at an assessment of the overall strength of evidence on the role and impact of private schools in developing countries (see Day Ashley et al. 2014).

Quality	Size	Context	Consistency
Strong: >50% of studies rated strong (with remainder of studies rated medium)	Strong: >10	**Strong** (5+ countries)	**Strong**: Findings are highly consistent, with >75% of studies clearly supporting or refuting assumption
Medium: ≤50% studies rated strong **(with remainder of studies rated medium)**	Medium: 6-10	**Medium** (3-4 countries)	**Medium**: Findings are moderately consistent, with 51% to 75% of studies clearly supporting or refuting assumption
Not used in the study (No low-quality studies included in the review)	Weak: ≤5	**Weak** (1-2 countries)	**Weak**: Findings are inconsistent, with ≤50% studies supporting/refuting assumption, or with a majority of neutral findings

Based on DFID's How to Note: Assessing the Strength of Evidence (DFID 2013).

This summary was then reflected in the corresponding evidence brief to visually demonstrate the consistency of the evidence and to demonstrate to readers how to visually identify the numbers of studies that were positive, neutral or negative with respect to a particular assumption and the strength of each individual study.

	[H1] QUALITY **Private schools are better than state schools**		**[H2] Equity** **Private schools provide education to disadvantaged children**	
	(A1) Private school pupils achieve better learning outcomes than state school pupils	(A2) Teaching is better in private	(A3) Private schools geographically reach the poor	(A4) Private schools are equally accessed by boys and girls
ASSESSMENT	[MODERATE +]	[STRONG +]	[WEAK O]	[MODERATE -]
Positive	India [15, 18, 20, 31, 35*, 41, 48] Kenya [11, 16] Nigeria [55] Pakistan [3, 6, 29*] Nepal [54]	India [15, 32, 33, 34*, 35*, 47*, 48, 55] Tanzania [26] Pakistan [3, 7] Nigeria [55] South Africa [45]	India [33, 42] Kenya [56]	India [50] Pakistan [3]
Neutral	Ghana [1] India [14, 21, 30, 47*, 58]	India [21]	Pakistan [3] South Africa [45] India [8, 59]	India [30, 41] Pakistan [17]
Negative	Kenya [39]	Kenya [39]	India [41]	Tanzania [26] India [23, 25, 34*, 42] Pakistan [6] Kenya [36*]

Summary evidence 1: Supply

Key:
STRONG = Body of evidence rated as 'strong' overall
MODERATE = Body of evidence rated as 'moderate' strength overall.
WEAK = Body of evidence rated as 'weak' overall.

\+ = Positive findings supporting assumption

\- = Negative findings refuting assumption

O = Neutral findings ambiguous in relation to assumption

***** = Numbered study assessed as high quality (remaining are medium)

	[H1] QUALITY Private schools are better than state schools		[H2] Equity Private schools provide education to disadvantaged children	
	(A1) Private school pupils achieve better learning outcomes than state school pupils	(A2) Teaching is better in private schools	(A3) Private schools geographically reach the poor	(A4) Private schools are equally accessed by boys and girls
ASSESSMENT	[MODERATE +]	[STRONG +]	[WEAK O]	[MODERATE -]
Positive	India [15, 18, 20, 31, 35*, 41, 48] Kenya [11, 16] Nigeria [55] Pakistan [3, 6, 29*] Nepal [54]	India [15, 32, 33, 34*, 35*, 47*, 48, 55] Tanzania [26] Pakistan [3, 7] Nigeria [55] South Africa [45]	India [33, 42] Kenya [56]	India [50] Pakistan [3]
Neutral	Ghana [1] India [14, 21, 30, 47*, 58]	India [21]	Pakistan [3] South Africa [45] India [8, 59]	India [30, 41] Pakistan [17]
Negative	Kenya [39]	Kenya [39]	India [41]	Tanzania [26] India [23, 25, 34*, 42] Pakistan [6] Kenya [36*]

Summary evidence 1: Supply

Key:
STRONG = Body of evidence rated as 'strong' overall
MODERATE = Body of evidence rated as 'moderate' strength overall.
WEAK = Body of evidence rated as 'weak' overall.

+ = Positive findings supporting assumption

- = Negative findings refuting assumption

O = Neutral findings ambiguous in relation to assumption

***** = Numbered study assessed as high quality (remaining are medium)

E. Evaluating the Overall Strength of a Body of Evidence

The following section presents a framework for assessing the strength of a body of evidence. The assessment framework for both single studies and bodies of evidence could be converted into a numerical calculator, though such an approach is not taken here.

Assessment of the overall strength of a *body* of evidence with reference to a particular policy or business case, is directly linked to the quality, size, consistency and context of the body of evidence. Where staff within the organisation are not able to assess all the individual studies that constitute a body of evidence due to inadequate time or expertise, they should (a) seek to use evidence synthesis products which *have* assessed the quality of individual studies, (b) commission evidence synthesis products which assess the quality of individual studies or (c) seek to make a judgement about a body of evidence based on the criteria outlined above.

Five categories are proposed to determine the overall strength of a body of research when it is being applied to a particular policy or programme design:

Table 1: Evaluating the Overall Strength of a Body of Evidence

Categories of Evidence	Combinations of Quality + Size + Consistency + Context	Typical Features of the Body of Evidence	What It Means
Very Strong	High quality body of evidence, large in size, consistent, closely matched to the specific context of the programme design/policy	The body of evidence includes studies based on experimental designs (including impact evaluations), as well as systematic reviews and/or meta-analysis.[21]	We are very confident that the intervention/research has the effect/associations anticipated or does not have the anticipated impact/associations. The body of evidence has few or no deficiencies. We believe that the findings are convincing and stable.

[21] Meta-analysis is used to refer to 'the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies that typify our attempt to make sense of the rapidly expanding research literature.' Glass, G.V., Primary, secondary and meta-analysis of research. *Educational Researcher*, *5*(10), 1976, pp. 5-8.

Categories of Evidence	Combinations of Quality + Size + Consistency + Context	Typical Features of the Body of Evidence	What It Means
Strong	High quality body of evidence, large or medium in size, generally consistent, matched to the specific context of the programme design/policy	The body of evidence is likely to include either experimental or quasi-experimental designs (including use of RCTs and statistical methods enabling causal identification). Non-experimental research designs (including comparative case study methods) that make attempts at counterfactual analysis are also likely to feature in these bodies of evidence, as are systematic reviews.	We are confident that the intervention/research has the effect/associations anticipated or does not have the anticipated impact/associations. The body of evidence has few deficiencies.
Medium	Moderate quality studies, medium size evidence body, generally consistent, which may or may not be relevant to the specific context of theprogramme design/policy; also covers limited number of high quality studies	The body of evidence is likely to include studies from multiple designs (qualitative and quantitative) but which have been assessed as being only of a moderate quality. The findings of the studies do not offer robust findings that can be derived and replicated across a range of contexts.	We are moderately confident that the intervention/research has the effect/associations anticipated or does not have the anticipated impact/associations. The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.
Limited	Moderate or low quality studies, small or medium size body, inconsistent, not matched to specific context of the programme design/policy	The body of evidence is comprised of studies based on varied designs and methodologies which do not meet the minimum standards of research quality. It includes causal inference derived from single case studies in a limited number of contexts and cross-sectional analysis performed in the absence of rigorous baseline data.	We have limited confidence that the intervention/research has/does not have the anticipated effect/associations. The body of evidence has major and/or numerous deficiencies. Additional evidence is needed to conclude that the findings are stable or that the intervention has the indicated effect.

Categories of Evidence	Combinations of Quality + Size + Consistency + Context	Typical Features of the Body of Evidence	What It Means
No evidence	No studies or impact evaluations exist		We have evidence of need but no evidence that the intervention/research does or does not have the effect/association indicated.

It is not realistic to expect all categories of evidence to attain a 'strong' or 'very strong' rating, especially where there is a nascent field or discipline with a limited number of studies. In such cases 'medium' will often be the best achievable rating and will be good enough.[22]

[22] This is also the conclusion of a review of grading systems in health research, which recognised that a high rating is not attainable for some disciplines. See Harbour, R. and J. Miller, A new system for grading recommendations in evidence based guidelines, *BMJ*, 2001, 323: pp. 334-6.

Appendix A: Table A1: Summary of Research Design and Methods

Type of Research: Primary and Empirical Research	Data Analysis Methods	Appropriate Use/Point of Application	Examples from Education Sector of High Quality Studies Using Methods
Observational (OBS)	Quantitative or qualitative	When inferring cause and effect *or* interpreting why something has happened.	
Cross-sectional data analysis /large-*n* surveys	Quantitative	When researchers wish to use inferential statistics to observe spatial variation and infer *cause and effect* relationships.	ASLAM, M., and G. KINGDON. What can teachers do to raise pupil achievement*? Economics of education review*, 2012, 30, 559-574 ASADULLAH, M.N., N. CHAUDHURY and A. DAR. *Assessing the performance of madrasas in rural Bangladesh*. Washington, DC: World Bank, 2009, pp. 137-48. KREMER, M., and K. MURALIDHARAN, K. Public and private schools in rural India. In P. PETERSON and R. CHAKRABARTI, eds., *School choice international*. Cambridge, MA: MIT Press, 2008.
Cohort/longitudinal analysis/panel data	Quantitative	To use inferential statistics to observe spatial variation and infer cause and effect relationships.	SANDEFUR, J. *On the evolution of the firm size distribution in an African economy*. Working Paper Series, CSAE-WPS-2010-5. Oxford: Centre for the Study of African Economies, 2010. BIRCHLER, K., and K. MICHAELOWA. *Making aid work for education in developing countries: an analysis of aid effectiveness for primary education coverage and quality*. WIDER Working Paper 2013/21. Helsinki: UNU-WIDER, 2013. GALAB, S., U. VENNAM, A. KOMANDURI, L. BENNY and A. GEORGIADIS. *The impact of parental aspirations on private school enrolment: evidence from Andhra Pradesh, India*. Young Lives Working Paper Series WP 97. Oxford: Young Lives, 2013.

Type of Research: Primary and Empirical Research	Data Analysis Methods	Appropriate Use/Point of Application	Examples from Education Sector of High Quality Studies Using Methods
Interviews/focus groups	Qualitative, descriptive analysis; thematic analysis; key events analysis	To gain deeper insight into people and communities. Not used to answer cause and effect, but may be useful in showing the *process* involved in causal relationships.	SRIPRAKASH, A. Pedagogies for development: the politics and practice of child-centred education in India. Sydney, Australia: Springer, 2012. GITHITHO-MURIITHIi, A. Education for all and child labour in Kenya: a conflict of capabilities? *Procedia - social and behavioral sciences*, 2010, 2(2), 4613-4621.
Ethnographic research	Qualitative, descriptive analysis; thematic analysis; key events analysis	To provide rich insights into people's views and actions. Emphasis on exploring the nature of social phenomena rather than testing any hypotheses.	CHUTA, N., and G. CRIVELLO. Towards a 'bright future': young people overcoming poverty and risk in two Ethiopian communities. Young Lives Working Paper No. 107. Oxford: Young Lives, 2013. ALTINYELKEN, H.K. Pedagogical renewal in sub-Saharan Africa: the case of Uganda. *Comparative education*, 2010, 46(2), 151-171. SARANGAPANI, P.M. Constructing school knowledge: an ethnography of learning in an Indian village. New Delhi: Sage Publications, 2003.
Case study research	Qualitative and/or quantitative.	To gain deeper understanding of subjects that offer revealing and interesting insights. Facilitates exploration of a phenomenon through a variety of lenses to explore and understand multiple facets of the phenomenon. Helps answer 'how' and 'why' questions.	ABD-KADIR, J., and F. HARDMAN. The discourse of whole class teaching: a comparative study of Kenyan and Nigerian primary English lessons. *Language and education*, 2007, 21(1), 1-15. SEBATANE, E.M., C. CHABANE and J.P. LEFOKA. *Teaching and lLearning in Lesotho: an empirical perspective of primary school classroom.* Ottawa: International Development Research Centre, 1992. MOLOI, F., N. MOROBE and J. URWICK. Free but inaccessible primary education: a critique of the pedagogy of English and mathematics in Lesotho. *International journal of educational development*, 2008, 28(5), 612-621.

Type of Research: Primary and Empirical Research	Data Analysis Methods	Appropriate Use/Point of Application	Examples from Education Sector of High Quality Studies Using Methods
Political economy analysis	Qualitative (could be supplemented by quantitative in a mixed-methods design)	Highly specialised analysis used to deepen understanding of the political context in a country to strengthen donor programming and design.	BEURAN, M., G. RABALLAND and K. KAPOOR. *Political economy studies: are they actionable? some lessons from Zambia.* Policy Research Working Paper 5656. Washington, DC: World Bank, 2011. WILKINSON, E. *Transforming disaster risk management: a political economy approach.* ODI Background Note Series. London: Overseas Development Institute, 2012.
Mixed-methods	Quantitative +qualitative:	To enable the researcher wishes to tackle a given research question from several relevant angles and/or more than one type of investigative perspective. Allows answering 'how', 'why', 'what' and 'where' questions.	ROELEN, K. and M. CAMFIELD. *A mixed-method taxonomy of child poverty: a case study from rural Ethiopia*. Young Lives Working Paper No. 76. Oxford: Young Lives, 2012. ORKIN, K. Are work and school complementary or competitive for children in rural Ethiopia? A Mixed-methods Study. Young Lives Working Paper No. 77. Oxford: Young Lives, 2012. HÄRMÄ, J. Low cost private schooling in India: is it pro poor and equitable? *International journal of educational development*, 2011, 31(4), 350-356.

Type of Research: Primary and Empirical Research	Data Analysis Methods	Appropriate Use/Point of Application	Examples from Education Sector of High Quality Studies Using Methods
Quasi-experimental (QEX)	Quantitative (can be supplemented by qualitative data)	To infer cause and effect. Is usually able to demonstrate presence and size of causal linkages with a reasonable degree of confidence.	
Propensity score matching (PSM)	Quantitative	When treatments cannot be 'randomised', this method serves as a powerful impact evaluation tool. Hinges on the notion that observed characteristics determine participation.	GODTLAND, E., E. SADOULET, A. JANVRY, R. MURGAI and O. ORTIZ. The impact of farmer–field–schools on knowledge and productivity: a study of potato farmers in the Peruvian Andes. *Economic development and cultural change*, 2004, 52 (1): 129-58. JALAN, J. and M. Ravallion. Estimating the benefit incidence of an anti-poverty program by propensity score matching. Journal of business and economic statistics, 2003, 21(1), 19-30.
Double difference methods (DD)	Quantitative (supplemented by qualitative in more recent designs)	To infer cause and effect relationships and evaluate impact (ex post) of programmes. Unobserved factors impact participation, but they do not vary over time and can therefore be 'differenced' out.	KHANDKER, S.R., Z. BAKHT. and G.B. KOOLWAL. The poverty impacts of rural roads: evidence from Bangladesh. *Economic development and cultural change*, 2009, 57(4), 685-722. CHAUDHURY, N. and D. PARAJULI. Conditional cash transfers and female schooling: the impact of the female school stipend program on public school enrollments in Punjab, Pakistan. Policy Research Working Paper 4102. Washington, DC: World Bank, 2006.

Type of Research: Primary and Empirical Research	Data Analysis Methods	Appropriate Use/Point of Application	Examples from Education Sector of High Quality Studies Using Methods
Regression discontinuity designs/Natural experiments	Quantitative	Can be used to determine cause and effect in the context of developing countries with limited data to carry out impact evaluations. Key disadvantage: Based on assessing very narrow thresholds (to ensure cases are similar).	DUFLO, E. Grandmothers and granddaughters: old age pension and intra-household allocation in South Africa. *World Bank economic review*, 2003, 17(1), 1-26. GALASSO, E. and M. RAVALLION. Social protection in a crisis: Argentina's plan jefes y jefas. *World Bank economic review*, 2004, 18(3), 367-400.
EXPERIMENTAL (EXP)			
Randomised control trials (RCTS)/ randomised designs/interventions	Quantitative – descriptive statistics and inferential statistics	To assess cause and effect relationships and demonstrate presence and size of causal linkages with a high degree of confidence. Able to create a robust counterfactual ('what if'). Allow answering 'who', 'what', 'where' and 'how' questions.	BERRY, J., D. KARLAN and M. PRADHAN. Social or financial: what to focus on in youth financial literacy training? Draft working paper. New Haven, CT: Innovations for Poverty Action, 2013. DUFLO, E., P. DUPAS and M. KREMER. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *American economic review*, 2011, 101(5): 1739-74. MURALIDHARAN, K., and V. SUNDARARAMAN. Contract teachers: experimental evidence from India. Working Paper No. 19440. Cambridge, MA: National Bureau of Economic Research, 2013.

Appendix B: Additional Resources on Assessing Evidence

There is already a range of resources and materials valuable for (a) strengthening individual capacity to assess strength of evidence and (b) appraising evidence when writing summary papers.

General guidance

a. EPPI centre resources – guidance and support on conducting systematic reviews
b. The Campbell Collaboration – systematic review guidance and support as well as other resources to researchers in the health and education sectors
c. Louise Shaxson's approach to evidence assessment for policy makers
d. International Institute for Environment and Development: 'Towards excellence: policy and action research for sustainable development'
e. Resources available on the ODI website, such as 'Evidence-Based Policymaking: What is it? How does it work? What relevance for developing countries?'
f. Information and guidance also available on 'What Works Clearinghouse Procedures and Standards Handbook'
g. Nancy Cartwright's work on evidence-based policy – a useful primer on the need to be careful when assessing evidence and applying evidence to different contexts

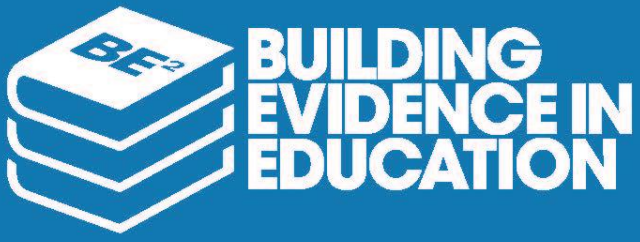
Evidence assessment frameworks

h. The GRADE approach to assessing quality of health research studies
i. The NICE Guideline Development Methods on assessing quality of health research studies
j. Critical Appraisal Skills Programme – multiple checklists for research quality of multiple research methods
k. Civil Service 'Rapid evidence assessment' framework from the HMG Government Social Research Unit – provides guidance relating to assessment of bodies of evidence

DFID Guidance

l.
a. DFID Broadening the range of designs and methods for impact evaluations, Report of a study commissioned by DFID (2012), working paper no. 38
b. Government Chief Scientific Adviser's guidance on the use of science and engineering advice in policymaking
c. DFID guide to research designs & methods
d. DFID Evaluation Handbook: guidance on evaluation designs & methods[23]
e. DFID Using Statistics How to Note
f. Use of Evidence in Policy-Making (Civil Service Learning online, forthcoming, Autumn 2014)

[23] See DFID Evaluation Department's handbook, chapter 4, Choosing your evaluation approach (design and methodology'. In addition, as of March 2014, DFID Evaluation Department was developing specific guidance on expected standards for the generation and use of strong qualitative research in evaluations.


BE²
BUILDING
EVIDENCE IN
EDUCATION