Assessing the Strength of Evidence

Contents

Introduction

Background: research and evidence in DFID

1. As an integral part of its work, the Department for International Development commissions research for a range of purposes:

   a. Research into the development of **new technologies or products** (e.g. new medicines for healthcare, new crop varieties for agriculture);

   b. Research to help understand **what kinds of development intervention are likely to work, and in what context** (e.g. the effects of increased government transparency on state accountability, or the effects of reducing trade barriers on investment and growth rates);

   c. Research to **strengthen understanding of the diverse political, social, economic and cultural contexts** in which DFID and its development partners operate (e.g. analysis of the economic opportunities available to women in specific countries, assessment of political or conflict dynamics in fragile states).

Why does the strength of evidence matter?

2. Research and evaluation generates the evidence required by public officials and civil servants to make informed judgements about how to design and implement policy, and how to spend scarce financial resources. Consequently, the identification and use of robust research and evaluation is integral to the Value for Money cycle.

What is the purpose of this guidance note?

3. Assessing the strength of evidence is a challenging task, and requires the application of technical knowledge and individual judgement. This How to Note[1] aims to help staff use evidence more judiciously for the benefit of designing and implementing effective policy and programmes. It introduces:

   i. the appraisal of the quality of **individual studies**;
   ii. the assessment of the strength of **bodies of evidence**.

4. More specifically, the Note (a) helps staff understand the distinctions between different types of research and what they can and cannot conclude on the basis of the research;

[1] This Note has been significantlymodified following extensive feedback on the initial version produced in February 2013.

(b) sets out common language that can be used in the discussion of the strength of evidence.

Scope and coverage of this Note

5. This Note sets out a number of principles which can be helpfully applied to assess the quality of single research studies, and recognise the characteristics of strong bodies of evidence. It draws from several academic disciplines, and from several sources of existing guidance.[2] It is relevant for the consideration of evidence generated by all research designs and disciplines (experimental, observational, quantitative, qualitative, natural and social science)[3].

6. The Note is **explicit in its recognition that different research designs and methods are more or less appropriate for answering different research questions**. Some research designs are essential for developing new technologies, others are valuable for answering 'what works' questions, whilst others are better suited to building rich contextual understanding.

A note on terminology

7. The terms 'quality', 'size', 'context', 'consistency' and 'strength' of evidence have specific meanings in this Note. Other approaches to assessing research evidence may use similar terms in a slightly different way.[4] This Note assumes that the overall 'strength' of a body of evidence is determined by the quality (or "avoidance of bias") of studies that constitute it, and by the size, context and consistency of the body of evidence.

---

[2] Examples include Spencer, L., et al., 2003, *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*, National Centre for Social Research, Cabinet Office, London. http://www.civilservice.gov.uk/wp-content/uploads/2011/09/Quality-in-qualitative-evaluation_tcm6-38739.pdf. Also see Government Social Research Service on Rapid Evidence Assessments.

[3] The ESRC 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. Standards of evidence are most developed in the health field. For health, the Cochrane Collaboration and Campbell Collaboration have established clear metrics for assessing research evidence and the conduct of systematic reviews. There is also a high degree of consensus on the basis for determining the quality of research evidence in the economics field. See http://www.thecochranelibrary.com; http://www.campbellcollaboration.org/library; GSDRC Helpdesk Research Report, *Qualitative Evaluation and Research*, 24 March 2012.

[4] What is termed 'strength of evidence' in the current note is typically referred to as 'quality of evidence' by the GRADE approach to assessing the quality of research. What is termed 'quality of research studies' in the current note is typically referred to as 'risk of bias' in the GRADE approach.

Applying this guidance note

8. The current Note is endorsed by DFID's Chief Scientific Adviser and Chief Economist. It is recommended reading for all DFID staff, and has the following applications.

DFID Evidence Products:

a. This Note is to be issued **as a guide** prior to the production of all synthesised evidence products. These include include Systematic Reviews, Evidence Papers, and Literature Reviews. It should be **comprehensively applied in the production of DFID Evidence Papers**. Whilst its formal application to other knowledge services products and commissioned research is discretionary, the Note serves as an indication of DFID's expectations for **all** discussions of research and the strength of evidence.

Business Cases:

b. Claims about the strength of evidence that feature in Business Cases, Ministerial submissions and policy papers should draw on this Note.

Box 1: Additional Resources

There are 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

i. Government Chief Scientific Adviser's Guidance on the Use of Science & Engineering Advice in Policy-Making;
ii. Use of Evidence in Policy-Making (Civil Service Learning online, forthcoming, Autumn 2014);
iii. Louise Shaxson's approach to evidence assessment for policy makers;
iv. International Institute for Environment and Development: 'Towards Excellence: policy and action research for sustainable development.'

Evidence assessment frameworks

v. The GRADE approach to assessing quality of health research studies;
vi. The NICE Guideline Development Methods on assessing quality of health research studies;
vii. Critical Appraisal Skills Programme: multiple checklists for research quality of multiple research methods;
viii. Civil Service 'Rapid Evidence Assessment' framework from the HMG Government Social Research Unit which provides guidance relating to assessment of bodies of evidence.

Available only for DFID Staff

ix. DFID Guide to Research Designs & Methods
x. DFID Evaluation Handbook: guidance on evaluation designs & methods[5]
xi. DFID Using Statistics How to Note.

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

Part I: Describing a single study

9. The current note recommends that single studies be described and categorised as follows:

   i. by type
   ii. by design
   iii. by method.

Type of research

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

    i. *Primary research studies* empirically observe a phenomenon at first hand, collecting, analysing or presenting 'raw' data.

    ii. *Secondary review studies* interrogate primary research studies, summarising and interrogating their data and findings.

    iii. *Theoretical or conceptual studies*: most studies (primary and secondary) include some discussion of theory, but some focus almost exclusively on the construction of new theories rather than generating, or synthesising empirical data.

Research Designs, Research Methods

Introduction

11. A research design is a framework in which a research study is undertaken. It employs one or more research methods to:

    i. collect data
    ii. analyse data.

12. **Data collection** can be either quantitative or qualitative.

13. **Data analysis methods** can also be quantitative (using mathematical techniques to illustrate data or explore causal relationships) or qualitative (collating 'rich' data and inferring meaning).

14. The line between quantitative and qualitative research is blurred by mixed method designs. Mixed methods may involve the quantitative analysis of qualitative data or the interrogation of quantitative data through a qualitative lens.[6] In that sense, different research designs and methods can be 'nested' as part of a flexible methodological approach to a research question.

15. Some designs are better suited for ***demonstrating*** the presence of a causal relationship, others are more appropriate for ***explaining*** such causal relationships while some designs are more useful for ***describing*** political, social and environmental contexts.

16. **Primary** research studies tend to employ one of the following research designs. As noted above, they may employ more than one research method.

    i. *Experimental* research designs (also called 'intervention designs', 'randomized designs' and Randomised Control Trials [RCTs]) have **two key features**. First, they manipulate an independent variable (for example, the researchers administer a treatment, like giving a drug to a person, or fertilizing crops in a field). Second, and crucially, they randomly assign subjects to treatment groups (also called intervention groups) and to control groups. Depending on the group to which the subject is randomly assigned, they will/will not get the treatment.

    The two key features of experimental studies increase the chances that any effect recorded after the administration of the treatment is a direct result of that treatment (and not as a result of pre-existing differences between the subjects who did/did not receive it). Experimental research designs use quantitative analysis (often 'descriptive statistics' followed by 'inferential statistics'). The combination of random assignment and quantitative analysis enables the construction of a robust 'counter-factual' argument (i.e. "this is what would have happened in the absence of the intervention or treatment"). Such designs are useful for demonstrating the presence, and size of causal linkages (e.g. "*a* causes *b*") with a high degree of confidence.

    ii. *Quasi-Experimental* research designs[7] typically include **one, but not both of the key features of an experimental design**. A quasi-experiment might involve the manipulation of an independent variable (e.g. the administration of a drug to a group of patients), but participants will not be randomly assigned to treatment or control groups. In the second type of quasi-experiment, it is the manipulation of the independent variable that is absent. For example, researchers might seek to explore the impact of the awards of scholarships on student attainment, but it would be unethical to deliberately manipulate such an intervention. Instead, the researchers exploit other naturally occurring features of the subject groups to control for (i.e.

---

[6] Stern, E. and others (2012). "Broadening the Range of Designs and Methods for Impact Evaluations." Department for International Development, Working Paper 38, p. 30.

[7] See California State University: Department of Psychology. 'Quasi-experiments.' Available at: http://psych.csufresno.edu/psy144/Content/Design/Nonexperimental/quasi.html

eliminate) differences between subjects in the study (i.e. they 'simulate' randomisation). A regression-discontinuity design is an example of a quasi-experiment.

iii. *Observational* (sometimes called 'non-experimental') research designs display **neither of the key features of experimental designs**. 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 or control groups. Instead, the researchers is merely an observer of a particular action, activity or phenomena. There are a range of methods that can be deployed within observational research designs:

- A variety of observational methods use **quantitative data collection and data analysis techniques** to to infer causal relationships between phenomena: for example, cohort and/or longitudinal designs; case control designs; cross-sectional designs (supplemented by quantitative data analysis) and large-n surveys are all types of observational research.

- Interviews, focus groups, case studies, historical analyses, ethnographies, political economy analysis are also all forms of observational research design, usually relying more on **qualitative methods** to gain rich understanding of the perspectives of people and communities.[8] When such studies are underpinned by structured design frameworks that enable their repetition in multiple contexts, they can form a powerful basis for **comparative research**.

Box 2: Research designs and methods as a means of reducing bias

If research is all about the quest for 'answers', then the consumers of research (whether they are policy-makers or community members) 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: users of research often have an appetite for patterns and predictability, and are curious to know if initiating action 'x' will result in consequence 'y'.

Some of the research designs and methods described explicitly seek to demonstrate **cause and effect** relationships, and are able to do so with varying degrees of confidence. Because they construct a 'counter-factual', experimental studies significantly reduce the risks of important biases affecting the findings of research, and for this reason, they are often regarded as the 'gold standard' for research which aims to isolate cause and effect.

However, research is not just about identifying cause and effect: it is also about understanding **why** some events unfold as they do, and learning more about why people have particular perspectives and interpretations of the events that affect them. This is often where the rich variety of observational (especially qualitative) research designs and methods add substantial value.

[8] Stern, E. et al. (2012). "Broadening the Range of Designs and Methods for Impact Evaluations." Department for International Development, Working Paper 38.

17. **Secondary review studies** tend to employ one of the following research designs:

i. *Systematic Review* designs adopt exhaustive, systematic methods to search for literature on a given topic. They interrogate multiple databases and search bibliographies for references. They screen the studies identified for relevance, appraise for quality (on the basis of the research design, methods and the rigour with which these were applied), and synthesise the findings using formal quantitative or qualitative methods. DFID Systematic Reviews are always labelled as such. They represent a robust, high quality technique for evidence synthesis. Even Systematic Reviews must demonstrate that they have compared 'like with like' studies.

ii. *Non-Systematic Review* 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.

iii. *Theoretical or conceptual* research studies may adopt structured designs and methods, but they 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'. Nor should existing policy papers or institutional literature.

Box 3: Research designs and methods: which is best?

This Note has already explained how some research designs and methods seek to address typical forms of bias in research. Some academic disciplines explicitly consider designs and methods hierarchically according to their relative ability to eliminate biases.[9]

However, DFID's work covers a huge span of economic, political and social policy and programmes. Given the range of purposes for which research is required in international development, this Note is clear that there **is no universally applicable hierarchy of research designs and methods**.

Instead, the Note argues that different designs are more or less appropriate for different research questions.[10] Indeed, some of the most powerful and evidence is produced when a range of methods are either 'mixed' together or used independently of one another (i.e. 'nested' within a broader methodological approach) to allow triangulation of findings. Typically, stronger bodies of evidence are likely to be characterised by the availability of a wide spectrum of evidence which uses, and triangulates findings from several research designs and methods.

[9] See for example, 'Levels of Evidence' diagram, Evidence-Based Practice in the Health Sciences, Evidence Based Nursing Tutorial.

[10] Stern, E. et al. (2012), p. 2. For a helpful overview of the different sorts of questions which are best answered by different research designs and methods, see Petticrew, M. & H. Roberts (2003), "Evidence, hierarchies and typologies: horses for courses." *Journal of Epidemiology and Community Health, 57*: 527-529, and Sandbrook, C. (2013), "Biodiversity, Ecosystem Services and Poverty Alleviation: What constitutes good evidence? A discussion paper." The Poverty and Conservation Learning Group Discussion Paper No. 10.

Why categorise studies by type, design & method?

18. The different types of study, different designs and methods outlined above are more or less appropriate for answering different types of research question. Categorising studies by type provides the reader with an initial, general understanding of how the study's findings were produced, and helps them begin to make some general judgements about the appropriateness of the design for the research question.

19. This Note recommends the use of the following descriptors to describe single research studies by type:

Research Type	Research Design
Primary (P)	Experimental (EXP) + state method used
	Quasi-Experimental (QEX) + state method
	Observational (OBS) + state method used
Secondary (S)	Systematic Review (SR)
	Other Review (OR)
Theoretical or Conceptual (TC)	N/A

How it looks in practice

20. In practice, synthesising evidence using this convention results in summaries of single studies as follows:

    i. For example, when citing a primary and empirical study by Jones, who uses an experimental research design, the citation may be written as (Jones, 2005 [P; EXP]).

    ii. In the case of an observational case study by Smith, the citation may be written as (Smith, 2004 [P; OBS, case study]).

    iii. 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]).

21. This Note strongly recommends that the method (not just the design) on which a single study is based should also be noted when it is cited.

Part II: Assessing the quality of single studies

22. Following the description of a single, primary research study by type, design and method the reviewer or user should aim to consider its quality (or the degree to which it

minimises the **risk of bias**). Although this is not a trivial exercise, there are some general rules of thumb that all staff will be able to apply.

Box 4: A note on quality assessment of secondary review studies

The current Note focuses principally on the quality assessment of **primary research studies**. The quality assessment of secondary reviews requires the use of a different set of criteria. Some standard questions to ask when assessing the credibility of secondary reviews include:

- 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.

For further guidance about what high quality secondary reviews look like, see Hagen-Zanker, J. & Mallett, R. (2013). 'How to do a rigorous, evidence-focussed literature review in international development.'[11]

23. When assessing the credibility of a study, 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 business case is an important, but separate, part of evidence synthesis, which is covered later in this How to Note.

Proxies for quality: journal rankings

24. Journal ranking systems can provide an indicative, though contested proxy guide to the scrutiny with which an academic study has been subjected prior to publication. The principal journal ranking system is the 'Impact Factor' rating. Journals often publish their Impact Factor ranking somewhere on their website.[12] An alternative is the 'H-Index',[13] which ranks individual academics according to productivity and impact.

25. 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 from southern academic organisations or those that feature in online journals, so a broad and inclusive approach is required to capture all relevant studies. Both 'Impact Factor' and 'H-index' scores may give an incomplete picture of academic quality.

---

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

[12] See, for example, the Quarterly Journal of Economics. A list of the highest impact journals classified by discipline is also available via Science Gateway, now hosted by Thomson Reuters.

[13] Hirsch, J. (2005). 'An Index to Quantify an Individual's Scientific Research Output.' Proceedings of the National Academy of Sciences, *102* (46). Available at: http://www.pnas.org/content/102/46/16569.full

Principles of high quality research studies

26. The following principles of credible research enquiry are relevant to all research studies. Reviewers of any research literature will have to think carefully about how exactly to apply these principles depending on the nature of the study. Assessors of studies will always have to make a judgement about study quality based on a combination of the following criteria. It is taken as read that research has been ethically conducted.[14]

i. **Conceptual framing**: high quality studies acknowledge existing research or theory. They make clear how their analysis sits within the context of existing work. They typically construct a conceptual or theoretical framework, which sets out their major assumptions, and describes how they think about the issue at hand. High quality studies pose specific research questions and may investigate specific hypotheses.

ii. **Transparency**: High quality studies are transparent about the design and methods that they employ, the data that has been gathered and analysed, and the location/geography in which that data was gathered. This allows for the study results to be reproduced by other researchers, or modified with alternative formulations. Failure to disclose the data and code on which analysis is based raises major questions over the credibility of the research. Transparency includes openness about any funding behind a study: research conducted with support from a party with vested interests (e.g. a drug company) may be less credible than that conducted independently.

iii. **Appropriateness**

There are three main types of research design (see above), and many types of methods. Some designs and methods are more appropriate for some types of research exercise than others.

Typically, experimental research designs tend to be more appropriate for identifying, with confidence, the presence of causal linkages between observable phenomena.[15] The implementation of an experimental design is not, in itself, a sign of good quality. The diverse array of observational (or 'non-experimental' designs) may be more appropriate for questions that either cannot be explored through experimental designs due to ethical or practical considerations, or for the investigation of perspectives, people and behaviours that lie at the heart of most development processes.[16]

iv. **Cultural sensitivity:**

Even research designs that appear well-suited to answering the question at hand may generate findings that are not credible if they fail to consider local, cultural factors that

[14] Criteria drawn up with reference to: Bryman, A. (2012). *Social Research Methods*. 4th Edition. Oxford University Press: Oxford.
[15] Stern, E. et al. (2012). "Broadening the Range of Designs and Methods for Impact Evaluations." Department for International Development, Working Paper 38.
[16] Ibid. pp. 18, 24.

might affect any behaviours and trends observed. For example, take a study that investigates efforts to boost girls' enrolment rates at schools in a religiously conservative country. If the study fails to explicitly consider the socio-cultural factors which influence parental support for girls' education, it is likely to miss the real reasons why the intervention worked or didn't work. High quality studies will demonstrate that they have taken adequate steps to consider the effect of local cultural dynamics on their research, or on a development intervention.

v. **Validity**: There are four principal types of validity.

Measurement validity: Many studies seek to measure something: e.g. agricultural productivity, climate change, health. Measurement validity relates to whether or not the specific indicator chosen to measure a concept is well suited to measuring it. For example, is income a valid measure of family welfare, or are specific measures of individual health and happiness more appropriate? Identifying valid measures is especially challenging and important in international development research: just because an indicator is a valid measure in one country or region does not mean it will be equally valid in another.

Internal validity: Some research is concerned with exploring the effect of one (independent) variable on another (dependent) variable. It can do so using a range of designs and methods. As described above, some designs and methods (e.g. experiments and quasi-experiments) are better able than others to determine such cause and effect linkages: they will minimise the possibility that some 'confounding', unseen variable is affecting changes in the dependent variable, and consequently they are said to demonstrate strong internal validity. Take the example of a study that explores the relationship between levels of corruption and firm efficiency. An internally valid study would employ a technique capable of demonstrating that corruption does indeed cause firms to become more inefficient. A study lacking in internal validity, on the other hand, might employs a technique which leaves open the possibility of reverse causality: i.e. that a firm is actually more likely to engage in corrupt behaviours because it is inefficient, and to compensate for its inefficiency.

External validity: This describes the extent to which the findings of a study are likely to be replicable across multiple contexts. Do they apply only to the subjects investigated during this particular study, or are they likely to apply to a wider population/country group? Quantitative researchers typically seek to address issues of external validity by constructing 'representative samples' (i.e. groups of subjects that are representative of a wider community/society).

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, and to do so without the research itself somehow impacting upon the subjects it seeks to study. Any time a researcher carries out an investigation in the field (asking questions, measuring something), s/he introduces something artificial into that context. Ecolocically valid studies will explicitly consider how far the research findings may have been biased by

the activity of doing research itself. Such consideration is sometimes referred to as 'reflexivity.'[17]

vi. **Reliability**: Three types of reliability are explored here.

Stability: if validity is about measuring the right 'thing', then stability is about measuring it 'right'. Assume that a study seeks to investigate the health of newborn children. Assume that 'birth weight' is a valid measure. For birth weight to be measured reliably, the investigator will require a reliable instrument (e.g. accurate weighing scales) with which to gather data. Alternatively, consider data which is gathered on the basis of questionnaires or interviews being conducted by multiple researchers: what steps, if any, have been taken to ensure that the researchers are consistent in the way they ask questions and gather data?

Internal reliability: many concepts can be measured using multiple indicators, scales and indices. For example, corruption could be measured by recorded incidence of embezzlement from public sector organisations, *and* with the use of a corruption perceptions index. If very significant discrepancies exist between indicators (e.g. if a country appears to experience low levels of corruption when embezzlement is measured, but high levels of corruption when perceptions are explored), then the internal reliability of one or other of the measures is open to question. High quality research will consider such issues, with specific attention to whether or not particular measures are well-suited to the cultural context in which they are taken.

Analytical reliability: the findings of a research study are open to question if the application of a different analytical technique (or 'specification') to the same set of data produces dramatically different results.

vii. **Cogency**:
A high quality study will provide a clear, logical 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 signpost the reader through the different sections of the paper, and avoid making claims in their conclusions that are not clearly backed up by the data and findings.

High quality studies will also be self-critical, identifying limitations of the work, or exploring alternative interpretations of the analysis.

27. A really rigorous review of the evidence on a given topic should give due consideration to all seven of these aspects of study quality. It is possible to construct checklists, or scorecards to grade evidence based on these criteria, and it is expected that DFID Evidence Papers will do so.

[17] See, for example, International Institute for Environment and Development (2012), Towards Excellence: Policy and Action Research for Sustainable Development, London: IIED.

Table 1: Principles of Research Quality

Principles of quality	Associated questions
Conceptual framing	Does the study acknowledge existing research?
	Does the study construct a conceptual framework?
	Does the study pose a research question or outline a hypothesis?
Transparency	Does the study present or link to the raw data it analyses?
	What is the geography/context in which the study was conducted?
	Does the study declare sources of support/funding?
Appropriateness	Does the study identify a research design?
	Does the study identify a research method?
	Does the study demonstrate why the chosen design and method are well suited to the research question?
Cultural sensitivity	Does the study explicitly consider any context-specific cultural factors that may bias the analysis/findings?
Validity	To what extent does the study demonstrate measurement validity?
	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 are the measures used in the study stable?
	To what extent are the measures used in the study internally reliable?
	To what extent are the findings likely to be sensitive/changeable depending on the analytical technique used?
Cogency	Does the author 'signpost' the reader throughout?
	To what extent does the author consider the study's limitations and/or alternative interpreations of the analysis?
	Are the conclusions clearly based on the study's results?

28. The following descriptors should be used when assessing the quality of single research studies. Directional arrows may be used to signify quality in DFID Evidence Papers. 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
High	↑	Comprehensively addresses multiple principles of quality.
Moderate	→	Some deficiencies in attention to principles of quality.
Low	↓	Major defincienes in attention to principles of quality.

Principles in practice

29. To summarise quality of evidence succinctly, reviewers may wish to abbreviate their quality assessment by use of an arrow (see above). However, if they do so, they must be prepared to defend their assessment based on the quality criteria spelled out. 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. Where many studies are reviewed by different analysts, this is particularly important to enable inter-rater reliability (i.e. that assessments of studies are more or less stable across multiple reviewers).

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

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

32. 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 been isolated in the current study).

Part III: Summarising the main characteristics of a body of evidence

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

i. The (technical) **quality** of the studies constituting the body of evidence (or the degree to which risk of bias has been addressed);
ii. The **size** of the body of evidence;
iii. The **context** in which the evidence is set;
iv. The **consistency** of the findings produced by studies constituting the body of evidence.

34. This section of the How to Note is intended to help DFID staff form judgements about the strength of evidence when identifying, sifting and assessing studies for use in Business Cases and policy papers.

Quality of the studies constituting the body of evidence

35. The quality of a body of evidence is determined by the quality of the single studies that constitute it (see Part II, 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 in Part IV, below). For example, it is possible for a body of evidence to be small in size, but high in quality.

36. 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.

Quality of the body of evidence	Definition
High	Many/the large majority of single studies reviewed have been assessed as being of a high quality, demonstrating adherence to the principles of research quality.
Moderate	Of the single studies reviewed, approximately equal numbers are of a high, moderate and low quality, as assessed according to the principles of research quality.
Low	Many/the large majority of single studies reviewed have been assessed as being of low quality, showing significant deficiencies in adherence to the principles of quality.

Size of the body of evidence

37. 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, does not constitute a large body of evidence.

38. Making a judgement about the apparent size of a body of evidence is also somewhat dependent on the research question, research context and subject area. When considering multiple dimensions of a major topic (take malaria as an example) it is useful to record which aspects of that topic (e.g. symptoms and diagnosis; prevention through drugs; prevention by other means; treatment; eradication) 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 domain of a particular academic field.

39. 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)

Context of the body of evidence

40. The reviewers of single studies should have noted the geography/context in which a study was conducted. This enables those describing a body of evidence to be clear about the context of the evidence that they are quoting. This is particularly important given that in many development sciences and programmatic interventions, the findings of research may be context-specific.

41. When determining the applicability of evidence from one context to another, the reviewer or policy-maker 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 policy-maker. 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.

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

Context
Global
Context Specific

Consistency of the findings of studies constituting a body of evidence

43. 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'.

44. 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.

45. 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 (contested)	One or more study/studies directly refutes or contest the findings of another study or studies carried out in the same context or under the same conditions.
Mixed	Studies based on a variety of different designs or methods, applied in a range of contexts, have produced results that contrast with those of another study.

Recap: summarising the main characteristics of a body of evidence

46. When summarising or synthesising evidence reviewers should seek to make a comment on the quality, size, context and consistency of a body of evidence. The following conventions can be used:

    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."

    b. **Or** "There is a medium-sized (+ no. 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 mixed."

    c. **Or** "There is a small-sized (+ no. 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.

Part IV: Evaluating the overall strength of a body of evidence

47. The following section presents a framework for assessing the strength of a body of evidence. Both the assessment framework for single studies, and for bodies of evidence could be converted into a numerical calculator, though such an approach is not taken here. The assessment framework is likely to be particularly useful for **examining the strength of evidence relating to a particular programme intervention**.

48. 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 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.

49. Five categories are proposed to determine the overall strength of a body of research when it is being applied to a particular policy or Business Case. This table is not intended as a formulaic checklist, but instead as an indicative guide to the typical features of very strong, strong, medium and limited bodies of evidence.

Table 2: Evaluating the overall strength of a body of evidence

Categories of evidence	Quality + size + consistency + context	Typical features of the body of evidence	What it means for a proposed intervention
Very Strong	High quality body of evidence, large in size, consistent, and contextually relevant.	Research questions aimed at isolating cause and effect (i.e. what is happening) are answered using **high quality experimental and quasi-experimental research designs**, sufficient in number to have resulted in production of a systematic review or meta-analysis.[18] Research questions aimed at exploring meaning (i.e. why and how something is happening) are considered through an **array of structured qualitative observational research methods** directly addressing contextual issues.	We are very confident that the intervention does or does not have the effect anticipated. The body of evidence is very diverse and highly credible, with the findings convincing and stable.
Strong	High quality body of evidence, large or medium in size, highly or moderately consistent, and contextually relevant.	Research questions aimed at isolating cause and effect (i.e. what is happening) are answered using **high quality quasi-experimental research designs and/or quantitative observational studies**. They are sufficient in number to have resulted in the production of a systematic review or meta-analysis. Research questions aimed at exploring meaning (i.e. why and how something is happening) are considered through **an array of structured qualitative observational research methods** directly addressing contextual issues.	We are confident that the intervention does or does not have the effect anticipated. The body of evidence is diverse and credible, with the findings convincing and stable.
Medium	Moderate quality studies, medium size evidence body, moderate level of consistency. Studies may or may not be contextually relevant.	Research questions aimed at isolating cause and effect (i.e. what is happening) are answered using **moderate to high-quality quantitative observational designs**. Research questions aimed at exploring meaning (i.e. why and how something is happening) are considered through a **restricted range of qualitative observational research methods** addressing contextual issues.	We believe that the intervention may or may not have the effect anticipated. The body of evidence displays some significant shortcomings. There are reasons to think that contextual differences may unpredictably and substantially affect intervention outcomes.
Limited	Moderate-to-low quality studies, medium size evidence body, low levels of consistency. Studies may or may not be contextually relevant.	Research questions aimed at isolating cause and effect (i.e. what is happening) are answered using **moderate to low-quality quantitative observational studies**. Research questions aimed at exploring meaning (i.e. why and how something is happening) are considered through a **narrow range of qualitative observational research methods** addressing contextual issues.	We believe that the intervention may or may not have the effect anticipated. The body of evidence displays very significant shortcomings. There multiple are reasons to think that contextual differences may substantially affect intervention outcomes.
No evidence	No/few studies exist.	Neither cause and effect, nor meaning is seriously interrogated. Any available studies are of low quality, and are contextually irrelevant.	There is no plausible evidence that the intervention does/does not have the effect indicated.

[18] 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, 5-8.

50. 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.[19]

Part V: Using and applying this guidance

51. This Note was initially designed for use by DFID staff. It is deliberately designed to support DFID's cross-cadre competency "Collating, analysing and presenting evidence/research using statistical and wider analytical skills," one of only four cross-cadre competencies.

52. The Note has potential application across a range of UK government departments and organisations working in international development. It should be used in conjunction with other materials (see Box 1, above) to maximise the use of evidence and to ensure judgements on the strength of evidence are well founded and consistent.

[19] 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 Miller, J., "A new system for grading recommendations in evidence based guidelines", *BMJ*, 2001, 323: 334-6.