Assessing the Strength of Evidence

Contents

Introduction ..................................................................................................................................................2
  Why does the strength of evidence matter? .........................................................................................2
  What is the purpose of this guidance note? .........................................................................................2
  A note on terminology .......................................................................................................................3
  How should staff apply this guidance note to their work? ...............................................................4

Part I: Describing a single study ..........................................................................................................5
  Type of research ....................................................................................................................................5
  Research designs, research methods .................................................................................................6
  Why categorise studies by type, design and method? .......................................................................8
  How it looks in practice .......................................................................................................................9

Part II: Assessing the quality of single studies ..................................................................................9
  Proxies for quality: journal rankings and citation frequency ...........................................................10
  Principles of high quality studies ......................................................................................................10
  How it is used in practice ...................................................................................................................15

Part III: Summarising the main characteristics of a body of evidence ..........................................16
  Quality of the studies constituting the body of evidence .................................................................16
  Size of the body of evidence ............................................................................................................17
  Context of the body of evidence ........................................................................................................17
  Consistency of the findings of studies constituting a body of evidence ........................................18
  Recap: summarising the main characteristics of a body of evidence ..............................................18

Part IV: Evaluating the overall strength of a body of evidence .........................................................19

Part V: Using and applying this guidance ..........................................................................................21
Introduction

Why does the strength of evidence matter?

1. Strong evidence is of central importance in informing policy and programming decisions across UK government departments. Robust research and evaluation generates the evidence required to form judgements, deliberate options and make intelligent decisions about how to spend scarce financial resources on behalf of taxpayers. It is critical to the work of all DFID staff and Ministers, and especially for the presentation of appraisal options in Business Cases.\(^1\) It is vital that research evidence is evaluated in a fair and balanced way.

What is the purpose of this guidance note?

2. This Note provides a thorough introduction to:

   a. the appraisal of the quality of **individual studies**;

   b. the assessment of the strength of **bodies of evidence**.

3. The Note is an integral part of DFID's commitment to equipping staff with the skills and specialist advice to help them improve their use of evidence. Other resources include:


   b. The Evaluation Handbook, providing guidance on evaluation designs and methods;\(^3\)

   c. 'Using statistics' How to Note;\(^4\)

   d. A set of guidance materials about how to summarise research evidence.\(^5\)

4. Assessing the strength of evidence is a challenging task. This Note sets a high standard for DFID staff. It requires a combination of technical knowledge and individual judgement. It may also require consultation with research specialists within and outside DFID. Proper assessment of evidence will help staff use evidence responsibly and judiciously for the benefit of better policy and programmes. Specifically, this Note will:

---

\(^1\) See, for example, http://www.bis.gov.uk/go-science/science-in-government/strategy-and-guidance

\(^2\) See the 'Research Methods' guide pages on InSight.

\(^3\) See Evaluation Department's Handbook, ch. 4, 'Choosing your evaluation approach (design and methodology)'.


\(^5\) See the Evidence Synthesis Sourcebook, available from DFID's Evidence into Action Team. Contact: w-evans@dfid.gov.uk.
a. help staff to broadly understand the distinctions between different data collection and analytical methods and what they can and cannot conclude as a result;

b. establish a common language that can be used in the discussion of the strength of evidence.⁶

5. This guidance is applicable to all categories of research and evaluation evidence used by DFID staff, especially in the social sciences. It applies to evidence generated through both quantitative and qualitative research methods.⁷ It recognises that some academic disciplines, such as medicine, and the methodologies associated with them, have a stronger tradition of assessing quality of research than social science disciplines.⁸ The Note references alternative evidence grading frameworks accordingly. To ensure consistency, this Note also draws on approaches developed in other parts of Government to assess research and evaluation evidence.⁹

6. A summary of this Note is incorporated into the revised Business Case guidance for ease of reference.

A note on terminology

7. Note that the terms ‘quality’, ‘size’, ‘context’, ‘consistency’ and ‘strength’ of evidence should be used with care in accordance with the definitions in this How to Note. This Note assumes that the overall ‘strength’ of a body of evidence is determined by the quality, size, context and consistency of the body of evidence.

---

⁶ The Note will be reviewed after one year based on feedback from staff and academic experts.
⁷ 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.
⁸ 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.
How should staff apply this guidance note to their work?

8. The current Note has been endorsed by the DFID Research Committee. As such, it is expected that DFID staff will apply the Note to their work as follows:

**DFID Evidence Products:**

9. DFID produces or commissions the following evidence products. It does so in conjunction with its partners and the Professional Evidence and Applied Knowledge Services (PEAKS) facility:

   a. Systematic Reviews: produced externally
   b. Evidence Papers: produced internally by DFID
   c. Literature Reviews: may be produced internally or externally
   d. Rapid Reviews: may be produced internally or externally
   e. Annotated Bibliographies: typically produced externally
   f. Evidence Maps: typically produced externally

The current Note should be applied to these products as follows:

   g. Parts I, II and III (see Contents, above) must be applied to the discussion and citation of evidence in the ‘Evidence Paper’ category of evidence products;

   h. It is highly desirable for Literature Review evidence products to draw on Parts I, II and III;

   i. This Note is to be issued as a guide prior to the production of all evidence products. It serves as an indication for DFID’s aspirations for all discussions of evidence. Its formal application to Rapid Reviews, Annotated Bibliographies and Evidence Maps and other knowledge services products is discretionary.

**Business Cases:**

   j. Business Cases, Ministerial submissions and policy papers should draw on Part IV (see Contents, above) of the guidance. They will preferably draw upon evidence synthesis products that have themselves been written with reference to Parts I, II and III.

**Alternative sources**

10. Where professional advisory cadres feel that the current Note does not fully serve their needs, they may wish to draw on established mechanisms in particular disciplines for assessing the strength of evidence. However, where they do so, they should ensure that the same aspects of evidential strength covered in this Note feature in their analysis.

11. Other sources of information for grading single studies and evaluating bodies of research include the following:
a. The Research Excellence Framework (REF) 2014 assessment methodology;
b. The GRADE approach to assessing quality of health research studies;
c. The NICE Guideline Development Methods on assessing quality of health research studies;
d. Critical Appraisal Skills Programme: multiple checklists for assessing the application of particular research designs and methods, the quality of single studies and bodies of evidence;
e. Civil Service ‘Rapid Evidence Assessment’ framework from the HMG Government Social Research Unit which provides guidance relating to assessment of bodies of evidence;
f. DFID Insight Evidence & Resources pages, containing guidance on critical appraisal;
g. Louise Shaxson’s evidence assessment template for policy-makers.

12. This HTN is accompanied by a short summary, which serves as a reminder of the main sections of the full Note.

**Part I: Describing a single study**

13. The current note recommends that single studies be described and categorised as follows:
   a. by type
   b. by design
   c. by method.

14. The following sections of this note explain how.

**Type of research**

15. This note recommends the categorisation of research studies by overarching type as follows:
   a. *Primary, empirical research studies* observe a phenomenon at first hand, collecting, analysing or presenting ‘raw’ data.
   
   b. *Secondary research studies* review other studies, summarising and interrogating their data and findings.

   c. *Theoretical or conceptual studies*, like secondary research studies, draw on previous research, but they do so primarily to construct new theories rather than generate, or synthesise empirical ‘evidence’.
16. A research design is a framework in which a research study is undertaken. It employs one or more research methods to:
   a. collect data
   b. analyse data.

17. Collected data can be either quantitative (data aggregated by numbers) or qualitative in nature.

18. Data analysis methods can also be quantitative (using numbers to illustrate data or demonstrate causal relationships) or qualitative (collating ‘rich’ data and inferring its meaning).

19. Increasingly, the line between quantitative and qualitative research is being blurred by the successful development of mixed method studies. Mixed methods may involve the quantitative analysis of qualitative data or the interrogation of quantitative data through a qualitative lens.10

20. Many research designs aim in some way to explore causal relationships. Some designs are better suited for demonstrating the presence of a causal relationship, whilst others are more appropriate for explaining 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.

21. **Primary & empirical** research studies tend to employ one of the following research designs, but as noted above, they may employ more than one research method.

   a. *Experimental* research designs (also called ‘intervention designs’ and ‘randomized designs’) administer a ‘treatment’ or ‘intervention’ to a ‘treatment group’, but not to a ‘control group.’ In such designs, the researcher deliberately manipulates the intervention (or ‘independent variable’) in order to explore its effects on the subject group. 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. Experimental research designs subject data exploring the subsequent behaviour of the two groups to quantitative analysis (specifically ‘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. “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. Randomised Control Trials (RCTs) are a well-established form of experimental research.

b. **Observational** (sometimes called ‘non-experimental’) research designs may be concerned with the study of groups that have received a ‘treatment’ with comparison groups that have not. However, unlike experimental research designs, it is not the researcher who deliberately manipulates the intervention: s/he is merely an ‘observer’ of a particular action, activity or phenomena (hence the name ‘observational’). Some ‘analytical’ observational studies use advanced quantitative analysis (specifically ‘inferential statistics’) to to infer causal relationships between phenomena. Others are more descriptive in nature, and may be more appropriate for teasing out explanations for causal relationships. The following are examples of observational research methods:

i. **Observational-analytical**: Cohort and/or longitudinal designs; case control designs; cross-sectional designs (supplemented by quantitative data analysis); large-n surveys.

ii. **Observational-descriptive**: Interviews, focus groups, case studies, historical analysis, ethnographies, political economy analysis.\(^\text{11}\)

---

**The risks of ‘selection bias’ and ‘confounding bias’**

A principal advantage of experimental designs (notably RCTs) over observational (also called ‘non-experimental’ designs) is that they reduce the risks of ‘selection bias’ and ‘confounding bias’. ‘Selection bias’ is the risk that the population receiving a treatment or intervention is somehow dissimilar to the ‘control’ group (a flaw which could invalidate a study). ‘Confounding bias’ is the risk that there is an additional, ‘unseen’ variable which is causing a particular effect. The ‘randomisation’ aspect of RCTs mitigates effectively for both of these risks, allowing the construction of a robust ‘counter-factual’ argument.

In some cases, specific sampling methods and ‘fixes’ can be applied to observational studies. Subjects are not assigned randomly to treatment or control groups, but patterns in the characteristics of the population are used by the researcher to control (partly) for selection bias and confounding bias. In such cases, studies are sometimes called **Quasi-Experimental designs**. Regression discontinuity design is an example of a quasi-experimental design.

---

22. The current Note avoids constructing a hierarchy of research designs and methods (though some disciplines do consider designs and methods hierarchically).\(^\text{12}\) It recognises that different designs are more or less appropriate to different contexts, and different

---


\(^{12}\) See for example, ‘Levels of Evidence’ [diagram](#), Evidence-Based Practice in the Health Sciences, Evidence Based Nursing Tutorial.
Counterfactuals are likely to be important for establishing the presence and strength of a causal relationship, but explanation for the nature of, and mechanisms behind causal relationships is often best achieved by observational designs using qualitative methods. Experimental and observational designs (both analytical and descriptive, quantitative and qualitative) can be used effectively together, with a range of methods either 'mixed' together. 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.

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

   a. **Systematic Review** designs adopt systematic methods to searching 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 and methods they employ), and synthesise the findings using formal quantitative or qualitative methods. Systematic Reviews are always clearly labelled as such. They represent a robust, high quality technique for evidence synthesis. Some caution should nevertheless be exercised in using the findings of Systematic Reviews: they must show that they are comparing 'like with like' studies. In addition, Systematic Reviews may suffer from issues relating to external validity (or 'generalisability' – see para 34 below for an explanation).

   a. **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.

   b. **Theoretical or conceptual** 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'.

**Why categorise studies by type, design & method?**

24. 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

---


14 http://dfidinsight/Other/Departments/EvidenceResources/Synthesizedevidenceproducts/Systematicreviews/index.htm
findings were arrived at, and helps the reader begin to make some general judgements about the credibility of the study.

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

<table>
<thead>
<tr>
<th>Research Type</th>
<th>Research Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary &amp; Empirical (P&amp;E)</td>
<td>Experimental (EXP) + state method used</td>
</tr>
<tr>
<td></td>
<td>Observational (OBS) + state method used</td>
</tr>
<tr>
<td>Secondary (S)</td>
<td>Systematic Review (SR)</td>
</tr>
<tr>
<td></td>
<td>Other Review (OR)</td>
</tr>
<tr>
<td>Theoretical or Conceptual (TC)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

How it looks in practice

26. In practice, synthesising evidence using this convention would result in summaries of single studies as follows:

   a. 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&E; EXP]).

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

   c. 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]).

27. 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

28. Following the description of a single study by type, design and method 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 all advisers will be able to apply. Staff may wish to consult colleagues who have particular expertise in a certain discipline, or Senior Research Fellows when faced with a particularly challenging technical study.

A note on assessment of secondary research studies

Note that assessment of study quality using these rules of thumb is possible for primary and empirical studies only. Systematic Reviews (when published officially as such) can be assumed to be of a high quality. The assessment of the quality of other, non-systematic reviews is a matter of judgement.
29. 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 and citation frequency**

30. 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 frequency with which a study or academic has been cited. The status of publications, in terms of the ‘impact factor’ of peer reviewed journals, can therefore inform an assessment of quality. DFID staff should treat academic peer-review as an important mechanism.

31. 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 in online journals, so a broad and inclusive approach is required to capture all relevant studies.

**Principles of high quality studies**

32. Whilst this Note 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 study should involve consideration of the relationship between the researcher and the subjects being studied and that appropriate ethical guidelines have been followed.

33. The first principle is a desirable feature of high quality studies:

   a. **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 lays bare their major assumptions. High quality studies pose specific research questions or hypotheses to which the research seeks to respond.

---

15 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 Index: http://www.highlycited.com/

34. The following principles are features **required** for a study to be considered of a high quality. They may be covered explicitly or implicitly by the author of a single study:

a. **Openness and transparency**: High quality studies should be transparent about the design and methods that have been employed, and the data that has been gathered and analysed. This allows for the study results to be reproduced by other researchers, or modified with alternative formulations. As such, failure to disclose the data and code on which analysis is based raises major questions over the credibility of the research. An important sign of quality is whether the author is being self-critical and open about limitations and alternative interpretations. Pointing out inconsistencies with other results helps put the study in its proper context. 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: ideally, a high quality study will be explicit about how it was funded.

b. **Appropriateness and rigour**:

   *What it means*: This refers to the appropriateness of the design and methods to the research question and its rigorous application. There are two main types of (see above), and many types of methods. 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 question than others.

   Typically, experimental research designs tend to be more appropriate for identifying, with confidence, the presence of causal linkages between observable phenomena. However, if the methods are improperly applied, it is possible for experimental studies to be of a low quality. The diverse array of observational (or ‘non-experimental’ designs) may be more appropriate for contexts and phenomena which cannot easily be explored through 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.\(^\text{17}\) Crucially, using an inappropriate method to tackle a question 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 for the study question and the context.

c. **Validity:**

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

**Measurement Validity:** During the data collection phase of a study, a researcher may set out to measure or interrogate a particular concept. S/he 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 wants to measure individual welfare, it has to choose 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. causing) 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 under-estimate the importance of a particular issue as a cause of an observed outcome or behaviour.

**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?

*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.

d. **Reliability:**

*What it means:* Reliability usually means one of two things. First, the reliability of a measurement means that not only is the right ‘thing’ being measured but also that it is being measured consistently and accurately. Second, the reliability of an analytical technique means that during the processing or use of data, the analysis is likely to produce consistent results when repeated multiple times.
An unreliable measurement instrument could potentially undermine an entire study. ‘Birth weight’ might be the right thing to measure in a piece of research, 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 was 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 for measuring 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.

e. **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. Try to consider whether or not you would have written the same conclusion or executive summary for the study based on the analysis and results it presents.

35. 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.
<table>
<thead>
<tr>
<th>Principles of quality</th>
<th>Associated principles</th>
<th>YES/NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual framing</td>
<td>Does the study acknowledge existing research?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the study construct a conceptual framework?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the study pose a research question?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the study outline a hypothesis?</td>
<td></td>
</tr>
<tr>
<td>Openness and transparency</td>
<td>Does the study present or link to the raw data it analyses?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the author recognise limitations/weaknesses in their work?</td>
<td></td>
</tr>
<tr>
<td>Appropriateness and rigour</td>
<td>Does the study identify a research design?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the study identify a research method?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the study demonstrate why the chosen design and method are good ways to explore the research question?</td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>Has the study demonstrated measurement validity?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is the study internally valid?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is the study externally valid?</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td>Has the study demonstrated measurement reliability?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Has the study demonstrated that its selected analytical technique is reliable?</td>
<td></td>
</tr>
<tr>
<td>Cogency</td>
<td>Does the author ‘signpost’ the reader throughout?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Are the conclusions clearly based on the study’s results?</td>
<td></td>
</tr>
</tbody>
</table>

36. 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 and Literature Reviews. 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.
<table>
<thead>
<tr>
<th>Study quality</th>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>↑</td>
<td>Demonstrates adherence to principles of appropriateness/rigour, validity and reliability; likely to demonstrate principles of conceptual framing, openness/transparency and cogency</td>
</tr>
<tr>
<td>Moderate</td>
<td>→</td>
<td>Some deficiencies in appropriateness/rigour, validity and/or reliability, or difficulty in determining these; may or may not demonstrate principles of conceptual framing, openness/transparency and cogency</td>
</tr>
<tr>
<td>Low</td>
<td>↓</td>
<td>Major and/or numerous deficiencies in appropriateness/rigour, validity and reliability; may/may not demonstrate principles of conceptual framing, openness/transparency and cogency</td>
</tr>
</tbody>
</table>

**How it is used in practice**

37. 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.

38. Returning to the previous examples, if a user of evidence cites a primary and empirical 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&E; EXP; →]).

39. 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.

40. 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).

41. Assessment of the quality of single studies should appear when cited in Evidence Papers. It is also good practice to follow this convention when drafting or commissioning other evidence products, and in Business Cases, provided that single studies have been subjected to critical appraisal. The assessment of the quality of single studies is a constituent part of summarising the overall value of a body of evidence which is commonly used in DFID Business Cases.
Part III: Summarising the main characteristics of a body of evidence

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

   a. The (technical) **quality** of the studies constituting the body of evidence;
   b. The **size** of the body of evidence;
   c. The **context** in which the evidence is set;
   d. The **consistency** of the findings produced by studies constituting the body of evidence.

43. 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

44. 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.

45. 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 is the case when assessing the quality of single research studies, but without needing to use directional arrows:

<table>
<thead>
<tr>
<th>Quality of the body of evidence</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Many/the large majority of single studies reviewed have been assessed as being of a high quality, demonstrating adherence to the principles of rigour, validity and reliability.</td>
</tr>
<tr>
<td>Moderate</td>
<td>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.</td>
</tr>
<tr>
<td>Low</td>
<td>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 rigour, validity and reliability.</td>
</tr>
</tbody>
</table>
Size of the body of evidence

46. 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.

47. 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 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 field.

48. 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.

<table>
<thead>
<tr>
<th>Size of body of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large (+ state number of studies)</td>
</tr>
<tr>
<td>Medium (+ state number of studies)</td>
</tr>
<tr>
<td>Small (+ state number of studies)</td>
</tr>
</tbody>
</table>

Context of the body of evidence

49. 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 is particularly important given that in many development sciences and programmatic interventions, the findings of research may be context-specific.

50. 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 DFID 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.

51. The descriptors of the size of the body of evidence are as follows:
Consistency of the findings of studies constituting a body of evidence

52. 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’.

53. 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.

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

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent</td>
<td>A range of studies point to identical, or similar conclusions</td>
</tr>
<tr>
<td>Inconsistent (mixed)</td>
<td>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.</td>
</tr>
</tbody>
</table>

Recap: summarising the main characteristics of a body of evidence

55. When summarising or synthesising evidence reviewers 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, s/he 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.”

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 inconsistent (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

56. 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.

57. 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.

58. 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:
<table>
<thead>
<tr>
<th>Categories of evidence</th>
<th>Combinations of quality + size + consistency + context</th>
<th>Typical features of the body of evidence</th>
<th>What it means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Strong</td>
<td>High quality body of evidence, large in size, consistent, and closely matched to the specific context of the business case.</td>
<td>The body of evidence includes studies based on experimental designs (including impact evaluations), as well as systematic reviews and/or meta-analysis.</td>
<td>We are very confident that the intervention has the effect anticipated or does not have the anticipated impact. The body of evidence has few or no deficiencies. We believe that the findings are convincing and stable.</td>
</tr>
<tr>
<td>Strong</td>
<td>High quality body of evidence, large or medium in size, generally consistent, and matched to the specific context of the business case.</td>
<td>The body of evidence is likely to include either experimental or quasi-experimental designs (including use of RCTs and statistical methods enabling causal identification). Observational 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.</td>
<td>We are confident that the intervention has the effect anticipated or does not have the anticipated impact. The body of evidence has few deficiencies.</td>
</tr>
<tr>
<td>Medium</td>
<td>Moderate quality studies, medium size evidence body, generally consistent, which may or may not be relevant to the specific context of the business case. Also covers limited number of high quality studies.</td>
<td>The body of evidence is likely to include studies from multiple designs (experimental and observational), 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.</td>
<td>We are moderately confident that the intervention has the effect anticipated or does not have the anticipated impact. The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.</td>
</tr>
<tr>
<td>Limited</td>
<td>Moderate or low quality studies, small or medium size body, inconsistent, not matched to specific context of the business case</td>
<td>The body of evidence is comprised of studies based on varied designs and methodologies, which do not meet the minimum standards of research quality. 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.</td>
<td>We have limited confidence that the intervention has/does not have the anticipated effect. Body of evidence has major and/or numerous deficiencies. Additional evidence needed to conclude that the findings are stable or that intervention has the indicated effect.</td>
</tr>
<tr>
<td>No evidence</td>
<td>No studies or impact evaluations exist</td>
<td></td>
<td>We have evidence of need but no evidence that the intervention does or does not have the effect indicated.</td>
</tr>
</tbody>
</table>

---

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.
59. 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

60. Staff should be strongly encouraged by managers to follow the approach set out in this How to Note when assessing the strength of evidence. It should be used in conjunction with other materials20 to maximise the use of evidence and to ensure judgements on the strength of evidence are well founded and consistent. Where possible, ThemeSites can be used to record and share appraisals of individual studies by staff.

61. There are implications for staff capacity in analysing, grading and using evidence as not all staff will have the same level of ability or expertise. ‘Collating, analysing and presenting evidence/research using statistical and wider analytical skills’ is one of four cross-cutting technical competencies for advisers which requires staff to be able to access, critically appraise and use evidence, demonstrating skills in the following areas:

   a. Understand a range of qualitative and quantitative research methodologies including the application of basic statistical methods;

   b. Critically appraise and assess the quality of published research and other potential sources of evidence;

   c. Interpretation, use and presentation of data and evidence in defining policy and practice.

62. Advisory staff will be expected to demonstrate basic technical competencies in these areas. RED will design and roll out the delivery of training modules in research and evidence to equip staff with the ability to analyse and interpret the strength of evidence in accordance with this How To Note.

---

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.

20 See the ‘Research Methods’ guide pages on InSight; Evaluation Department’s Handbook, ch. 4, ‘Choosing your evaluation approach (design and methodology)’; Global Statistics Partnerships Team, ‘Using Statistics’ How to Note, forthcoming; Evidence Broker’s Guide to Evidence Synthesis, available from DFID’s Evidence into Action Team. Contact: w-evans@dfid.gov.uk; matthew-harvey@dfid.gov.uk.