

O'Brien-Malone, A. & Diamond, M. R. (2009, September). *How should an evaluator make judgments about the quality of research evidence from the social sciences?* Paper presented at the 2009 Australasian Evaluation Society International Conference, Canberra, Australia.

How should an evaluator make judgments about the quality of research evidence from the social sciences?

Dr Angela O'Brien-Malone¹ & Dr Mark Diamond²

Abstract

The use of evidence in policy-making is seen as highly desirable — indeed, it is seen as 'best practice'. This means that policy makers need to be able to assess evidence, especially evidence drawn from the social sciences. But often the publications which are relevant to the policy question are vast in number, variable in quality, and inconsistent in conclusions. One of the central problems for evaluators is that they are often asked by their policy-maker clients to filter the evidence relevant to a policy intervention. This filtering of evidence happens in many circumstances – for example, when an evaluator is asked to contribute to a literature review, or to the choice of members of a consultative panel. Central to the evaluator's role as a filter of evidence is the ability to make judgments about the quality of research publications and to interpret the conclusions of these publications. A number of hierarchies have been developed to inform judgments about the quality of research publications, including hierarchies published by illustrious sources such as the UK Cabinet Office Social Exclusion Task Force. The hierarchies are based on the notion that much can be gained from the simple process of weighting more highly the conclusions of research papers which have used better methods. Though true, the notion is not as simple as it seems.

In this paper we examine a selection of published hierarchies, discuss some of their deficiencies, and identify a number of strategies that will overcome many of those deficiencies. We commend these strategies to evaluators.

¹ angela.obrienmalone@utas.edu.au

² www.markdiamond.com.au. Further discussion of related topics can be found on this web site.

Last year one of us (Angela) gave a paper at the AES Conference in Perth. That paper was called *Evidence and Policy Making: The Evaluator's Contribution*. At the end of the paper someone in the audience asked her *how* to assess evidence. This paper is our answer to that question.

When we put in the abstract earlier this year we were thinking largely about methodological issues, but we later realised that there are issues that are perhaps even more slippery than things to do with methods. So although we touch briefly on methodology, we deal mostly with the other issues.

There is a bunch of what might be termed 'rules' in circulation and these rules give some guidance for assessing the quality of scientific evidence. For instance, there are rules like parsimony: *Never posit two causes when one will do*.

But the admonishment to favour parsimony is probably more useful for scientists constructing scientific theories than it is for those who, like evaluators, have to assess how good the evidence for those theories is without themselves having extensive expertise in the domain.

Let's not mince words here: Assessing evidence is hard; it requires concentrated thought and judgement. As Andrew Leigh (www.andrewleigh.com) has said, the so-called 'soft sciences' should actually be called the hard sciences!

When we began thinking about this topic we did a search for information and guidance which might already exist and we found surprisingly little. Between us, we have 50 years of experience assessing different kinds of evidence. We are methodologists, so our expertise is in knowing how to know; and we frequently work with bodies of evidence related to topics where we are not experts *on the topic*. This is what evaluators must do when they are confronted with the need to assess evidence from the social sciences. So in thinking about this paper, what we did was try and work out what implicit rules and processes we use when faced with the task of assessing some new evidence. This paper is essentially a reflection on our own experience rather than an exhaustive coverage of the topic but it will provide a starting place for developing the topic.

In this paper we will present a series of examples about assessing evidence; examples which we have then summarised as nine "rules". We have, just because it amused us, called these rules Angela and Mark's Rules for Assessing Evidence — which has the acronym AMRAE, a word which in Irish means 'a wonder' or 'a marvel'!

Our first example is of things your mother might have told you when you were a child, at least if you are about our age. These things include:

- Swimming after lunch will make you drown;
- Drinking milk when you have a cold will make you develop more mucus.

Both these assertions are myths (see footnote for more information).³ There is nothing special about water and the contents of your stomach. Water and stomach contents do not

³ *Swimming after lunch will make you drown.*

- Debunking Summer Health Myths
<http://www.medicinenet.com/script/main/art.asp?articlekey=47368>
- American Red Cross — <http://www.nyredcross.org/page.php/prmID/418>
- About.com: Pediatrics — http://pediatrics.about.com/library/owt/bl_swimming_eating_r.htm

Drinking milk when you have a cold will make you develop more mucus.

- Mayo Clinic — <http://www.mayoclinic.com/health/phlegm/AN01455>
- Common Cold — a group of medical professionals and scientists
<http://www.commoncold.org/special1.htm>

mysteriously interact so that entering the water with food in your stomach (not even lunch food!) will immediately and irrevocably lead to your demise by drowning.

Similarly, milk protein is digested and broken down just like any other protein. It is not absorbed into your blood and exuded through your nose unchanged! This leads to our first rule.

Rule 1: Check for congruence with other knowledge

The first AMRAE is: Check for congruence with other things you know about the world. It turns out that this seemingly obvious rule is more difficult than it might first appear.

American Scientist earlier this year published a paper by Gelman and Weakliem in which they critique a recent series of biology papers with titles like *Violent men have more sons* and *Beautiful parents have more daughters*. The basic thesis of the papers is that

- being beautiful benefits girls more than boys
- being violent benefits boys more than girls
- consequently the male-to-female ratio in the children of beautiful or violent parents is biased away from the normal.

Now, there is a well-established literature on the variation in the sex ratio of human births under various conditions. Variations in sex ratio of up to 1 percent have been seen, for instance, in conditions of starvation or extreme deprivation. Also, selective abortion in some countries has raised the proportion of male births up to 1.4 percent higher than expected. The biology articles discussed by Gelman and Weakliem report a difference in sex ratio of 4.3 percent (and some media reports had this effect as high as 36 percent!). So here we have a claim that the beauty of the parents is biasing the sex ratio by four times the amount that selective abortion can. That is so unexpected, so contrary to prior evidence, that one should be very sceptical about the result. Yet, the biology papers were published in peer reviewed journals and, as Gelman and Weakliem said, “publication in a peer-reviewed journal seemed to have removed all skepticism” (p. 314).

It’s easy to decide that we wouldn’t be taken in by such papers, or to laugh at the urban myths our mothers believed. But maybe we harbour a few of our own. Here’s a bunch of assertions many of us will have seen or heard, and which we might believe to be true, but are they?

- Eating blessed thistle (*Cnicus benedictus*) can help nursing mothers produce more milk.
- Having your neck manipulated can improve your eyesight.
- For healthy living, an adult must drink a minimum of eight glasses of water a day.
- The brain is only plastic for certain kinds of information during specific ‘critical periods’, so the first three years of a child are decisive for later development and success in life.
- Exposure to ‘enriched environments’ in early childhood enhances the brain’s capacity for learning.
- There are visual, auditory and haptic learners.

Are these assertions correct? The answers are:

- We don’t know but the best we can work out is that medical science doesn’t know either
- Yes (but with caveats)
- No

- No
- No
- No

If you want a source for information on the last three questions, the OECD has a [web page](#) on Neuromyths⁴. Let us talk a little about the second assertion, regarding neck manipulation, because it leads into our second and third rules.

“Having your neck manipulated can improve your eyesight.” When we first came across this assertion we rejected it outright. We said to ourselves, ‘*What has your neck got to do with your eyes?*’ However, we were forced into revising our position when a good source told us that there was some scientific research which suggested that this was correct. We did some checking and, yes, at least in some circumstances, the changes in the retina brought about by migraines might be able to be moderated or reduced by neck manipulation, possibly because neck manipulation brings about changes in cerebral blood flow. Of course, neck manipulation also carries significant risks, including “death by chiropractor”! This leads us to Rules 2 and 3.

Rule 2: Be wary of your own intuitions

Rule 3: Check that the source is authoritative

Be wary of your own intuitions.

There are many things about human behaviour which are not “common sense”. There are examples of well-meaning programs based on common sense, which went terribly wrong — sometimes so wrong that they actually made the problem they set out to solve worse. One well-known example is the “Scared Straight” programs run with adolescents whose delinquent behaviour places them at risk of ending up in juvenile detention. The Scared Straight approach is based on the idea that if these juvenile delinquents visit a juvenile prison of some kind it will make them aware of how unpleasant it is and scare them out of their delinquent behaviours. Sounds like common sense. But juveniles who are put through these programs show a pattern of an *increase* in reoffending. Far from helping, the programs make things worse. Various explanations have been offered for this effect—but what is unmistakable is that the approach does not achieve the desired outcome.

Here is another example. Mandatory drug testing has operated in UK prisons since the 1990s. The idea is that random urine tests of prisoners are conducted and if their urine shows evidence of drug use, they get some additional punishment. It was expected that random testing of urine would act as a deterrent to the use of drugs in prisons. However, a pilot study conducted in 1995 showed that the use of opiates and tranquilizers increased with the introduction of the random testing program. Why? Possibly because some prisoners switched from using marijuana to using heroin—marijuana use remains detectable in urine for up to three weeks, while heroin use is only detectable for two or three days. So the result of this policy was an increase in the use of heroin and other opiates—which probably wasn’t the desired outcome. The size of the increase is the subject of debate, but that the program produced a shift in the kind of drugs used is now established.

Is the source authoritative?

Who are the leading people in the field? Whose papers are being cited most frequently? Who has had a long and consistently good research career? If you talk to people working in areas of research surrounding the one you are interested in, which names are consistently mentioned as being people of significance?

⁴ See http://www.oecd.org/document/4/0,3343,en_2649_35845581_33829892_1_1_1_1,00.html

Of course, positive answers to any or all of these questions do not guarantee the quality of a specific piece of work that a researcher has published. Every one can make a mistake. But the answers do tell you something about the general quality of a researcher's work. A similar set of criteria operate for sources and web-sites such as universities, government, and so on.

Of course what one really wants to know is how good the *specific* piece of research is. One way of trying to judge whether a paper is good is to think about the journal in which it is published. Good journals are likely to have published good work and so are authoritative sources. But how do we know whether a journal is good?

The impact factor of journals is the average number of times their published papers are cited and this can be used as a measure of journal quality. But journal rankings based on measures like impact factors are not without their problems. Journals which have a high impact ranking show a bias towards being English speaking, monied, and published in a Western country. Journal impact rankings also depend on the size of the citing audience. If there are only 1000 real experts in a highly specialist field—say, quantum computing—then probably the only people doing any citing of the published papers are these 1000 people. Journals designed for specialist audiences tend to have lower citation rates than journals which are generalist. But research published for a specialist audience may be of better quality than that published for a generalist audience. So while impact factors are a place to start, they are not the only thing to think about when making a judgement on the likely quality of a paper. While journals with a high impact factor are likely to be good, those with a low impact factor are not necessarily bad.

Finally, even highly reputable, generally excellent journals can occasionally make mistakes in what they choose to publish. Publication in a leading peer-reviewed journal does not automatically indicate that the research is good—it is simply a sensible place to start when making a judgement about quality.

Rule 4: Look for dissent and criticism.

We have already seen one example of published dissent and criticism in the Gelman and Weakliem paper on sex ratios. Another example comes from the Cochrane Review. The Cochrane Review was set up precisely because making judgements about evidence is hard. The Cochrane library is an on-line repository of reviews of evidence on health issues. Cochrane Reviews all use meta-analytic approaches and aim to evaluate issues based on only the highest quality evidence; they operate at a very high standard, and they rate as one of the most reliable sources of information on health matters. But even a Cochrane Review can come to a mistaken conclusion.

The Cochrane Review entry on using mammography to screen for breast cancer concluded that the risks incurred due to over-diagnosis and over-treatment based on false positives outweigh the benefits from that screening. We were convinced by the review but were later introduced to a paper by Freedman et al. Now David Freedman was arguably the greatest applied statistician of the second half of the 20th century, and the analysis he and his colleagues conducted on the original Cochrane Review concluded that the authors of the review had erred in the way that they selected papers for inclusion in their meta-analysis (i.e., the studies included in the review were systematically biased) and that the benefits associated with the use of mammographic screening for detecting breast cancer are substantially greater than the risks. We changed our opinion.

The lesson? Always look for authoritative criticism.

Rule 5: How much did it cost?

Rule 6: Where does the funding for the research come from?

How much did it cost?

How should the sheer volume of research papers on a topic be weighed? For instance, there are an awful lot of studies using functional magnetic resonance imaging (fMRI) which have been published over the last ten years. What does this imply? One thing it could mean (possibly does mean) is that fMRI studies are relatively cheap to do if you have access to the technology. The sheer volume of papers should be discounted as an indicator of anything, if the process of producing these papers is significantly cheaper than doing other kinds of research. The worst case scenario is if doing confirmatory research is cheap and easy, and doing disconfirmatory research is expensive and hard. Medieval ornithologists, for example, had a much easier time examining confirmatory evidence for the hypothesis that *All swans are white*. Gaining the disconfirmatory evidence was a lot harder... first one had to discover Terra Australis....

Where does the funding for the research come from?

Ask yourself whether funding sources stand to gain anything from the claimed results. Look for papers where the underpinning funding and power discrepancies at play are analysed.

One example of an area where honest answers to questions about who is funding what are currently very important is in the ongoing controversy surrounding the efficacy of antidepressants and the use of psychotropic medications in the treatment of various bipolar disorders. The background story involves pharmaceutical companies, universities and large funding bodies, and is far too long and complicated to try and summarise here. It is a story of lawsuits, possible corruption, vast amounts of money, difficulty accessing data, suicide and homicide; in short, this story has all the ingredients of a who-done-it with none of the escapist elements.

David Healy is a researcher in the field of psychotropic medications. Healy has done a lot of work investigating the psycho-social nature of the medical/clinical/diagnostic edifice surrounding particular psychological disorders, including depression and bipolar disorder. In general, Healy leans to the view that drug companies have, by dint of advertising, created an inflated estimation about the prevalence of particular psychological disorders in the minds of treating clinicians and in the minds of the public. Healy also contends that the drug companies have created an impression which overestimates the efficacy of their medications. Healy is not alone, and he has been published in journals as significant and reputable as PLoS Medicine, so it is fair to say that this is a serious controversy.

Disentangling the threads of this controversy is well beyond what can be covered today, but on the basis of authoritative sources (e.g., PLoS) one would have to decide to proceed with caution in any review of this area, and to exercise particular caution regarding the money trail.

Just to emphasise this point. On 25 July 2009, PLoS Medicine and the New York Times won a court battle to obtain documents which revealed industry tampering with medical literature. The game goes like this. You are a presumably 'independent' academic author. A ghost writer writes a paper for a medical journal and puts your name on it as the author. The ghost writer is in the pay of a drug company and puts the company's point of view. The ghost writer gets money, you might too. Following their court victory, fifteen hundred documents that contained details on how articles which used ghost writers to highlight specific marketing messages were placed in the medical literature were released to PLoS Med and the New York Times. It pays to watch the money trail.

Rule 7: Check that the research is methodologically sound.

A number of hierarchies have been developed to inform judgements about the quality of research publications, including hierarchies published by illustrious sources such as the UK Cabinet Office Social Exclusion Task Force. The hierarchies are based on the notion that much can be gained from the simple process of weighting more highly the conclusions of research that used better methods.

One thing of note is that many hierarchies rank meta-analytic studies as being of particular importance when evaluating evidence. For instance, the UK social Exclusion Taskforce writes of meta-analyses that they are the “[b]est source of assurance that an intervention works (or doesn’t).” The increasing number of recommendations to policy makers and others to give meta-analytic studies a high weight in their evaluation of evidence means that users of social scientific evidence must become adept at making judgements about the quality of meta-analyses.

Meta-analyses have already been mentioned several times in this paper, and you might be getting the idea that there are special problems to look out for if you are reading a meta-analysis; if so, then you’re right. Meta-analyses suffer from the same “garbage in—garbage out” problems as all forms of analysis. That is, they are only ever as good as the selection process which leads to the decision to include, or discard, studies from the analysis. Meta-analyses rest fundamentally and inescapably on a judgement about quality of research evidence, and that judgement is often quite opaque to the reader of the published meta-analysis. The previously mentioned Cochrane Review paper on mammography, and the subsequent criticism by Freedman et al., revolved around the selection of studies for inclusion in the analysis. If poor judgement is exercised about what studies to include (or not include) then the resulting meta-analysis is in the ‘garbage out’ category. This means that reading someone else’s meta-analysis doesn’t avoid the need for judgement on the part of the reader.

In addition to the judgement problem, there is also what is known as the ‘bottom drawer’ problem—an expression that refers to all those papers which did not get published and so got stored in the bottom drawer of someone’s filing cabinet. The politics of publication are such that getting a null result or a failure to replicate published is almost impossible. Rejection rates from most journals are upwards of 90%. Only the sexy papers get published and null results are not sexy. Journals are now often profit making enterprises. They want a readership, and (to quote Gelman and Weakliem) articles which have titles like: *There is no compelling evidence that beautiful parents are more or less likely to have daughters* just don’t cut the mustard. Authors of good meta-analyses will try and track down such bottom drawer papers by writing to likely researchers, or by trawling the web for PDFs stored in unusual places, and so on. But this is all a fishing trip and must necessarily be variable in its success. One must therefore expect that there will be a general tendency for meta-analyses to be biased towards finding a result and that findings of no-difference will be underrepresented in meta-analytic outcomes.

Rule 8: Extraordinary claims require extraordinary proof.

The initiating events in *The Broxtowe Case* occurred in the borough of Broxtowe in Nottinghamshire in the UK in 1987 where it was alleged that children were abused in the course of satanic rituals. As a consequence of the allegations, children were taken away from their families and made Wards of Court. While the children were in care, their foster parents were required to keep diaries recording anything which the children said or did which might be relevant. Two years later, ten adults appeared in court charged with 53 offences including incest, indecent assault and cruelty towards children. The adults were

found guilty and received long prison sentences. In the wake of this trial, other ritual abuse cases were prosecuted in other parts of the UK and many other children were taken away from their families. Some of the evidence against the adults included the diary entries made by the foster carers.

Subsequent to the trial a Joint Enquiry Team was set up. The report that the Team produced is a damning indictment of the investigation into the accusation of ritual abuse. None of the so-called evidence in the case was found to hold up under scrutiny. In particular, the diary records from the foster parents showed that the children had mentioned the following 'events' that had been used as evidence of abuse:

- a member of the family putting on a cloak and flying, and children being turned into frogs by the witches
- babies being killed by being jumped upon and then left in the garages
- babies being taken from next door and from across the road and having their heads bashed on the floor
- babies being thrown on the bonfire
- the family having dead babies hung around their necks
- killing people at witch parties
- all their family eating a man killed in a boat
- children being killed at parties
- being sexually abused in an underground room at the church, all the family and neighbours dressing up there as witches, clowns and ghosts
- underground rooms, tunnels, burning sticks, a bath full of blood, a red drink that makes you sleepy, a table with a purple plastic cloth with snakes, moons, stars, pigs, spiders and donkeys on it.

No evidence was found to corroborate these statements. No children were found to have been turned into frogs. None of the convicted adults appeared to have the ability to fly through the air. There were no tunnels, and there were no missing children or adults which fitted the bill.

The report of the Joint Enquiry concluded that there was no evidence of Satanic ritual abuse in the Broxtowe case and that it was the fact that parts of the Social Services Department had developed a belief in ritualistic or satanic abuse which had resulted in children being encouraged to believe in and allege weird abuse. Many similar "satanic abuse" and "ritual abuse" cases also appeared in the US during the 1980s and were ultimately found to be the stuff of fantasy. Of course, to some extent the damage was done: Children had been removed from their homes and adults had been wrongly imprisoned.

The claims made and the charges laid in the Broxtowe case were truly extraordinary and should have required extraordinary evidence. In fact, the evidence was non-existent.

Rule 9: Always look for alternative explanations.

Recently a paper was published in which it was asserted that text messaging using the predictive text function on mobile phones increased impulsiveness in children (Abramson et al., 2009). I will not go into detail about this paper; suffice it to say that the study has [a number of problems](#). The heart of this study is a task in which the children were asked to respond to words which were presented in a written form on a computer screen. The task was one where both the time to make a response and the number of errors were critical variables. Technically it was a reaction time task.

Now the thing about reaction time tasks is that the instructions which participants are given are very important, because speed in responding and accuracy in responding trade-off against one another, and the type of instructions given to participants bias that trade-off one way or another. If you are told your fingers will be cut off if you make an error you go very slowly to avoid mistakes. If you are told that you will get \$10 for every response which you make which is faster than 100 milliseconds you go very fast and you don't worry about accuracy. So you might have expected that something would have been said in the paper about the task instructions. But the paper is silent on this point.

So when we are told that the children who used text messages frequently responded more quickly and less accurately than other children we have to ask some questions. Does this result tell us something about impulsiveness? Not necessarily. If the children were instructed only to respond quickly and no mention was made of accuracy then it might be that the text-messaging children are simply better at doing as they are instructed. In the absence of clarity on this central point the paper is uninterpretable, and at least one alternative explanation can be offered for the results.

So without more details, and in the absence of controls for alternative explanations, the paper has little credibility, even though it has gained a lot of press coverage.

Summary

- Rule 1** **Check for congruence with other knowledge.**
- Rule 2** **Be wary of your own intuitions.**
- Rule 3** **Check that the source is authoritative.**
- Rule 4** **Look for dissent and criticism.**
- Rule 5** **How much did it cost?**
- Rule 6** **Where does funding for the research come from?**
- Rule 7** **Check that the research is methodologically sound.**
- Rule 8** **Extraordinary claims require extraordinary proof.**
- Rule 9** **Always look for alternative explanations.**

Which brings us to the uncomfortable punch line of this paper: There is no magic bullet. Assessing the quality of evidence requires making a judgement. The basis for that judgement should be reason and argument. We have, in the final analysis, no other tools upon which we can rely.

Bibliography and additional commentary

Abramson M.J., Benke G.P., Dimitriadis C., Inyang I.O., Sim M.R., Wolfe R.S., Croft R.J. (2009). Mobile telephone use is associated with changes in cognitive function in young adolescents. *Bioelectromagnetics*. Epub ahead of print.

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Banks, G. (2009). *Evidence-based policy-making: What is it? How do we get it?* Paper presented in the ANZSOG/ANU Public Lecture Series, Canberra, ACT.

- Freedman, D., Petitti, D. B., & Robins, J. M. (2004). On the efficacy of screening of screening for breast cancer. *International Journal of Epidemiology*, 33(1), 43–73.
- Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power. *American Scientist*, 97(4), 310–316.
- Gore, S. M., Bird, A. G., & Ross, A. J. (1996). Prison rights: mandatory drugs tests and performance indicators for prisons. *British Medical Journal*, 312:1411–1413.
- Gøtzsche, P. C., & Nielsen, M. (2005). Screening for breast cancer with mammography. *Cochrane Database of Systematic Reviews*, 2, Art No.: CD001877. (First published in 2001.)
- Healy, D. (2006). The latest mania: Selling bipolar disorder. *PLoS Med*, 3(4): e185.
- Leigh, A. (June 6, 2009). What evidence should social policymakers use? *Australian Treasury Economic Roundup*, 1, 27– 43, 2009. Available at <http://ssrn.com/abstract=1415462>
- Leigh, A. (17–18 August, 2009). Evidence-Based Policy: Summon the Randomistas? Paper presented at Strengthening *Evidence-Based Policy in the Australian Federation, Productivity Commission Roundtable*, Canberra, ACT.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- In a very famous study published by Nisbett and Wilson in 1977 (*Telling more than we can know: Verbal reports on mental processes*) the authors showed that, at least sometimes, when people make decisions they can report those decisions, but they have not a clue about what is influencing their judgement or decision. Nisbett and Wilson did things like set up a “consumer survey” in a shopping mall where they stopped passers by and asked them to give their opinion on which of a set of four identical pairs of ladies stockings were the best quality. Overwhelmingly (by a factor of almost 4:1) people chose the right most pair. Even more surprising, people were willing to offer all sorts of explanations for their decisions and judgements, but these explanations were completely awry. People said that the pair on the right was a little finer in weave; the colour was a bit more even; the sheen was a fraction more... No one, not one person, said that they chose the pair because it was on the right—and when the researchers suggested this explanation they all “denied it, usually with a worried glance at the interviewer suggesting that they felt either that they had misunderstood the question or were dealing with a madman.” The “common-sense” explanations of their own judgements which Nisbett and Wilson’s participants offered were wrong. We can be just as easily misled if we base our judgements about what other people will do on “common sense”.
- Singleton, N., Pendry, E., Simpson, T., Goddard, E., Farrell, M., Marsden, J., & Taylor, C. (2005). *The impact and effectiveness of mandatory drug testing in prisons*. Home Office.
- Social Exclusion Task Force (2008). *What Works? Guidance on evidence-informed commissioning and monitoring of services for vulnerable groups*. Appendix 3. Cabinet Office.
http://www.cabinetoffice.gov.uk/social_exclusion_task_force/think_research.aspx
 Retrieved 29 August 2009.