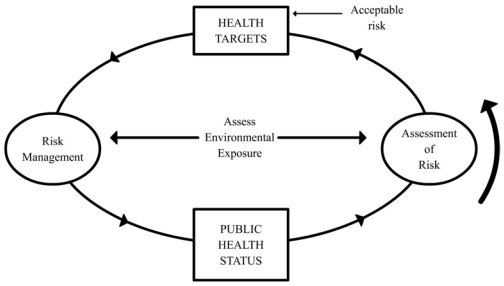

9

Quality audit and the assessment of waterborne risk

Sally Macgill, Lorna Fewtrell, James Chudley and David Kay

In order to avoid the 'garbage in, gospel out' scenario described by Burmaster and Anderson (1994) it is becoming increasingly clear that there is a need for some sort of standardised quality assessment to examine the strength of the inputs to the assessment of risk area. This chapter proposes one possible approach and notes the need for further development in this area. While the examples draw heavily on the risk assessment area, the same approach can be used for any of the tools driving the assessment of risk.

9.1 INTRODUCTION

How strong is the science for assessing waterborne health risks? Unless the answer to this question is known, then how can risk assessment or

 Water Quality: Guidelines, Standards and Health. Edited by Lorna Fewtrell and Jamie Bartram. Published by IWA Publishing, London, UK. ISBN: 1 900222 28 0

epidemiological study results be sensibly interpreted and acted upon? How can it be known with what degree of confidence, or of caution, to proceed?

These questions arise from the acknowledged limitations of science to provide definitive inputs to the assessment of waterborne risk. There are gaps and limitations in the current state of scientific knowledge. These are not identified here as limitations in competence or motivation of scientific experts. They are instead identified as intrinsic structural limitations in the fields of research which are being drawn upon.

Weinberg (1972) introduced the concept of trans-science to refer to problems which can be formulated within traditional scientific paradigms (for example as testable hypotheses) but which are beyond the capability of science definitively to resolve. Categories of problem falling within this realm include those entailing experimental set-ups that would be logistically too complex to co-ordinate in practice (owing to the sheer size and complexity of the technology or the sheer number – possibly millions – of experimental species required), problems raising ethical issues (notably the wrongs of experimentally exposing people to harmful substances), and problems where surrogate indicator species have to be studied in the absence of accessibility to true species or pathogens.

Other structural limitations, of a conceptually simpler nature, arise from the brute force of economics. Science is expensive, and it is simply not possible to fund all that would be desirable. For example, of the universe of toxic and carcinogenic chemicals that are as yet untested there is a fundamental issue in setting research priorities of whether it is better to test all of them less intensively, or intensively study a small proportion (Cranor 1995). At the same time, there are some problems that might be solved, if research priorities were such that the right team could be resourced to address itself to them. The interdisciplinary nature of some problems can of itself make them intrinsically less attractive for individual research funders to champion.

It is therefore possible to visualise a spectrum of risk assessment issues based on the strength of the available science in each case (Figure 9.1). Trans-scientific problems, by nature, lie towards the right of this range, classic laboratory science towards the left.

Figure 9.1. Spectrum of uncertainty.

For sensible interpretation of results, users of risk assessments and the studies that may feed into such risk assessments need to know where, along this spectrum, the science relevant to any particular issue lies. Put more strongly, as

consumers of the products of science, they need a 'charter' of the quality of what they are being given.

9.2 UNCERTAINTY IN ASSESSMENT OF WATERBORNE RISKS

'One problem with quoting quantitative predicted risks is that the degree of uncertainty is quickly forgotten.' (Gale 1998, p. 1)

Uncertainties in the assessment of waterborne risks will be identified here with reference to the general paradigm for risk assessment provided by the USA National Academy of Sciences. This presents risk characterisation (the core scientific process of estimating risk) as the integration of three distinct stages (NAS 1983).

(1) Hazard assessment looks at the nature and strength of evidence that an environmental agent can potentially cause harm. The evidence may come from tests on animals, coupled with inferences about possible human effects; or from case studies of people known to have been exposed to the agent of interest; or from human volunteer experiments. There are widely recognised limitations in extrapolating animal findings to human populations. There are difficulties in being absolutely sure that the observed responses are indeed caused by the suspected substance, and not by some other cause. There are doubts about how representative an experimental group is of a population more generally, or of sub-groups that may be particularly susceptible. There are differences in treatment efficiencies.
(2) Dose–response assessment aims to specify the relationship between the dose of a substance and the extent of any resulting health effects. Calibration of dose–response models may lead to the identification of critical threshold levels below which there are no observed adverse effects, or alternatively to representation of the classic U shape of the dose–response relationship for chemical essential elements (moderate doses beneficial to health; low and high doses both harmful to health). The conclusions from dose–response assessments are often

controversial, as there can be large measurement errors, misinterpretation of symptoms and often conclusions rely on statistical analysis which is vulnerable to misuse. It is particularly difficult, perhaps impossible, to specify a dose–response model for low levels of concentration. The translation of findings from one species to another as well as from one population to another is problematic.

(3) Exposure assessment seeks to establish the intensity, duration and frequency of the exposure experienced by a human population. There is a great deal of uncertainty here, owing to difficulties in measuring dilute concentrations of substances far from their originating source, limits of detection of some substances, and lack of specific knowledge about species recovery and viability. There are also problems in predicting population distribution patterns relative to those concentrations, in knowing water consumption rates, and in lack of awareness of specific local conditions (such as plumbing or hygiene conditions).

The overall risk characterisation, as the integration of these three stages, produces an estimate of the severity and likelihood of a defined impact resulting from exposure to a specified hazard. It is sometimes expressed as a number or a range. In more sophisticated studies, Monte Carlo analysis might be included as part of the approach (e.g. Medema et al. 1995), in order to account for the full distribution of exposure and dose–response relationships in a distribution of risk. This conveys information on the relative imprecision of the risk estimate, as well as measures of central tendency and extreme values (Burmaster and Anderson 1994). There is, however, no generally accepted way of conveying the overall strength of results (for example of the confidence that one can place in estimated probability distributions, which in turn depends on the state of the science and quality of the data utilised).

Taking Cryptosporidium in tap water as an example, authors have reported a variety of risk assessment results, summarised in Table 9.1. Haas and Rose (1994) have also calculated that during the Milwaukee cryptosporidiosis outbreak people would have been exposed to 1.2 (0.42–4.5) oocysts/litre to account for the level of illness seen.

Table 9.1. Risk assessment results – *Cryptosporidium* in tap water

Risk (95% CI)	Comments	Reference
9.3×10^{-4} ($3.9 \times 10^{-4} – 19 \times 10^{-4}$)	Daily risk of infection with drinking water containing 1 oocyst/10 litres	Rose *et al*. 1995
3.6×10^{-5} ($3.5 \times 10^{-7} – 1.8 \times 10^{-3}$)	Daily risk of infection associated with drinking water supplied from a conventional surface water treatment plant in the Netherlands	Medema *et al*. 1995
0.0009 (0.0003–0.0028)	Median annual risk of infection from exposure to 1 oocyst per 1000 litres of water in non-AIDS adults	Perz *et al*. 1998
0.0019 (0.0003–0.0130)	Median annual risk of infection from exposure to 1 oocyst per 1000 litres of water in AIDS adults	Perz *et al*. 1998
3.4×10^{-5} ($0.035 \times 10^{-5} – 21.9 \times 10^{-5}$)	Daily risk of infection from exposure to New York drinking water	Haas and Eisenberg 2001
0.0001	Annual acceptable risk of infection from drinking water	Macler and Regli 1993

9.3 THE CASE FOR QUALITY AUDIT (QA) OF SCIENCE IN RISK ESTIMATES

'Quantitative risk analyses produce numbers that, out of context, take on lives of their own, free of qualifiers, caveats and assumptions that created them.' (Whittemore1983, p. 31)

Limitations in science generate uncertainty in estimates of waterborne risk. As things currently stand, this uncertainty is of unknown (or unreported) extent and degree. It is without a generally accepted published measure. This is considered to be an unsatisfactory state of affairs that could in principle be addressed if some kind of quality audit was systematically practised. In order for this to be possible, appropriate audit tools need to be developed and tested.

Quality audit has become an increasingly familiar practice in many areas. The higher education sector in the UK, for example, is now familiar with the systematic quality auditing of research and of teaching activity in all university departments, and of the different methodologies that are used in each case. Other examples include the use of certification schemes in product labelling to reflect high quality standards, and more generally the various International Organisation for Standardisation (ISO) quality initiatives.

In the context of waterborne risk management problems, the corresponding need is to know about the quality, strength, or degree of certainty of the science underpinning the risk estimates. However, whereas in the case of teaching quality assessment a low score typically indicates remediable weaknesses ('could do better'), in the case of a quality audit of science, the weaknesses identified are not necessarily remediable.

If it is recognised that uncertainty is an intrinsic quality of many of the fields of science relevant to waterborne risk assessment, then objectively it should be a matter neither of shame nor of concealment to acknowledge this position. On the contrary, it should become a matter of standard practice faithfully to reflect significant uncertainties as part of the 'findings' about how big the risks really are. At the same time, given that it is as yet not standard practice, then research is needed to investigate the best way of doing this, ultimately to develop and refine an appropriate formal protocol for representing and communicating related aspects. Tentative examples of such protocols have begun to emerge in the literature. Further development and testing is needed as a foundation for wider promotion and acceptance of their principles.

If appropriate quality audit tools could be developed and applied, then this should benefit scientific communities by meeting the need for faithful representation of the strength of the knowledge base, thereby, for example, protecting academic disciplines against over-confidence in their outputs and pre-empting accusations of overselling. It should also provide intelligence for the management of research priorities according to areas of uncertainty which are most critical for contemporary policy issues.

Correspondingly it should benefit policy communities and other users of scientific outputs, by providing a diagnostic basis from which to facilitate interpretation of scientific inputs to environmental policy. This will guard against the conferment of undue authority on findings from inherently immature fields (for example, in the setting of regulatory standards). At the same time it should guard against unfounded criticism or rejection of more definitive results. It should also reduce conflict and promote more efficient decision-making, by proactively targeting particularly critical areas:

> To improve the validity of risk estimates, quality assurance principles should be rigidly implemented, and tools for this purpose should be developed. A particular point of attention is the development of structured, transparent methods to precipitate expert opinions in the risk assessment process. (Havelaar 1998)

9.4 A PROPOSED QUALITY AUDIT FRAMEWORK

In the absence of a generally accepted quality audit (QA) procedure for risk assessment science already in the literature, a pragmatic approach, starting from first principles, is presented below. This is based on a checklist of criteria against which the strength of scientific inputs to risk characterisation can be systematically evaluated. This will pinpoint the nature of weaknesses, and provide an overall view of the strength of risk estimates.

The composition of the checklist takes its inspiration from the work of Funtowicz and Ravetz (1990) who pioneered a new numerical symbolism (notation) for representing uncertain scientific inputs to policy decisions. In demonstrating a preference for a checklist as distinct from a notational system, however, what is given below deliberately departs from the Funtowicz and Ravetz formulation. The checklist approach is preferred because it is conceptually simpler while at the same time being systematic and offering flexibility.

The starting point is a simple conceptual representation of the process of producing scientific inputs for waterborne risk assessment (Figure 9.2).

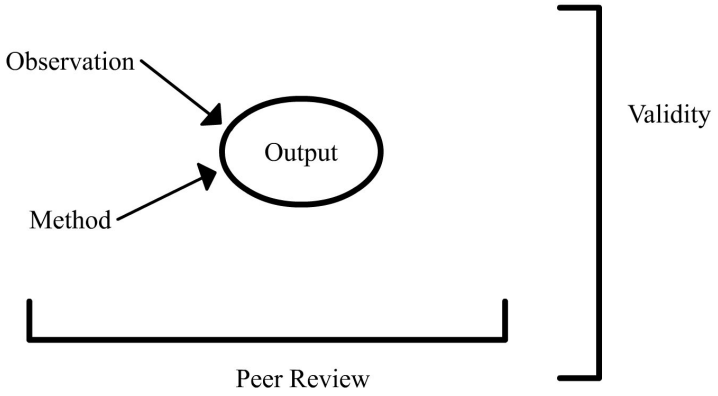

Figure 9.2. Conceptual representation of the quality audit framework components (reprinted from Macgill *et al*. 2000, with permission from Elsevier Science).

As with all scientific endeavour, this process has an empirical or observational aspect (data), and a theoretically informed methodological aspect. These two 'input' aspects combine to produce (as an 'output') an estimate of risk probability, risk magnitude or dose–response effects (or whatever) according to context. Given that the authority or standing of any such 'outputs'

can only ultimately be assessed following discourse and review among a peer community, each quantification process should, in principle, be subject to peer review. Consensus, on the basis of peer review, must be a necessary condition for producing definitive quantification. Finally, the relevance (or validity) of the quantified outputs to a particular context of interest must be accounted for.

Having established a conceptual model, each of its aspects provides grounds for interrogating the strength of the scientific inputs to waterborne risk assessment.

For the observational aspect we may ask:

- How close a match is there between the phenomenon being observed to provide data input, and the measure adopted to observe it?
- How reliable is the data or empirical content ?
- How critical is the data to the stability of the result?

For the theoretical/methodological aspect we may ask:

- How strong is the theoretical base?
- How resilient is the result to changes in theoretical specification?

For the result itself we may ask:

- Has a true representation of the real world been achieved?
- Is the degree of precision appropriate?

For the process as a whole we may ask:

- How widely reviewed has it been and what is the reviewers' verdict?
- What is the degree of consensus about the state of the art of the field?

And for the appropriateness to any particular applied context we may ask:

- How relevant is it to the intended application?
- To what extent can we be assured of its completeness?

These five categories of question provide the basis of a checklist for examining the quality of scientific inputs to assessments of risk (Table 9.2). The nature of the questions that have been identified within each category will be

considered more fully below, together with further background to the suggested scales for recording an evaluative response to each question.

Table 9.2. Outline quality audit framework

Dimension	Criterion	Question	Level	Score
Observation	Measure	How close a match is there between what is being observed and the measure adopted to observe it?	Primary	4
			Standard	3
			Convenience	2
			Symbolic	1
			Inertia	0
	Data	How strong is the empirical content?	Bespoke/Ideal	4
			Direct/good	3
			Calculated/limited	2
			Educated guess	1
			Uneducated guess	0
	Sensitivity	How critical is the measure to the stability of the result?	Strong	4
			Resilient	3
			Variable	2
			Weak	1
			Wild	0
Method	Theory	How strong is the theoretical base?	Laws	4
			Well-tested theories	3
			Emerging theories/comp models	2
			Hypothesis/stat processing	1
			Working definitions	0
	Robustness	How robust is the result to changes in methodological specification?	Strong	4
			Resilient	3
			Variable	2
			Weak	1
			Wild	0
Output	Accuracy	Has a true representation of the real world been achieved?	Absolute	4
			High	3
			Plausible	2
			Doubtful	1
			Poor	0
	Precision	Is the degree of precision adequate and appropriate?	Excellent	4
			Good	3
			Fair	2
			Spurious	1
			False/unknowable	0

Table 9.2 (cont'd)

Dimension	Criterion	Question	Level	Score
Peer review	Extent	How widely reviewed and accepted is the process and the outcome?	Wide and accepted	4
			Moderate and accepted	3
			Limited review and/or medium acceptance	2
			Little review and/or little acceptance	1
			No review and/or not accepted	0
	State of the art	What is the degree of peer consensus about the state of the art of the field?	Gold standard	4
			Good	3
			Competing schools	2
			Embryonic field	1
			No opinion	0
Validity	Relevance	How relevant is the result to the problem in hand?	Direct	4
			Indirect	3
			Bare	2
			Opportunist	1
			Spurious	0
	Completeness	How sure are we that the analysis is complete?	Full	4
			Majority	3
			Partial	2
			Little	1
			None	0

Also 'scores' under each criterion for unknown (–) and not applicable (n/a)

9.5 THE FIVE ASPECTS OF THE QA FRAMEWORK

9.5.1 Observation

Three types of potential empirical weakness have been identified: first, weaknesses in the appropriateness of the measure used to observe a given phenomenon of interest; second, weaknesses in the extent of empirical observation (data) available; third, sensitivity of results to changes in data inputs.

Weaknesses in the appropriateness of the measure used to observe a given phenomenon potentially arise because there is often no direct (fundamental) measure of the phenomena of interest, so an indirect measure has to be used. Well-known examples include: the use of indicator species; the spiking of laboratory samples to infer 'untraceable' elements; the use of sampling to infer characteristics of a larger (unobservable) population; the use of available (rather than desirable) levels of aggregation or resolution, for example, in measures of pollutant levels; the use of laboratory animals as 'surrogates' for human subjects; the tendency for census enumerators simply

to count what is obvious to their own common sense with no guarantee of consistency from one enumerator to another. In all such cases it is desirable to know how well the given indicator represents what it is being used to depict. A qualitative scale for representing this is included in Table 9.2. Corresponding scales are suggested for all other criteria below.

A good empirical base is a prerequisite for definitive science. However, in practice, and notably in the field of environmental risk assessment, the quality of data collection can be extremely variable. Considerations of cost, for example, may mean that water quality measurement is restricted to a single sample at a given site, rather than a range of samples at different depth and spatial co-ordinates across that site.

In principle it is possible to conceive a quality range running from reliable primary data of controlled laboratory standard, or as compiled by a first-rate task force, to secondary data of lesser quality – including proxy measures and sheer guesswork (educated or otherwise). While inexpert guesses will typically be given little if any standing, educated guesses should also be interpreted with caution, because of the potential of systematic biases.

The criterion of sensitivity asks whether results are resilient to changes in inputs (data, parameter values, etc.). Formal sensitivity analysis can test this to some extent, examining the existence and impact of critical values, and framing answers in explicit probability terms. Where sensitivity analysis has not been undertaken, one may wish to judge estimates rather differently from where it has.

9.5.2 Method

There would be little more than a ‘chaos of fact’ if there were no coherent recognition of why certain sorts of measurement were wanted, and not others, if no general patterns could be discerned among the different elements of empirical evidence available, if there were no awareness of what constituted critical measurement, or if there were no intelligent base to the way in which empirical inputs were to be processed or combined in a model. The theoretical aspect comes into play here.

Depending on the degree of understanding of the real world, this may range (at best) from laws to (less than desirably) working definitions. The hypothesis is the elementary testable theoretical statement for the study, which may be either refuted or accepted. Even the ‘emerging theory’ place on the scale has only a score of ‘2’, because of susceptibility to hypothesis errors.

Robustness calls for an examination of the resilience of the output (or estimate) to a change in theoretical specification. In some cases, change in

theoretical specification may have little effect, while in others, change in model specification may be critical.

9.5.3 Output

This aspect explores possible deficiencies arising from the formal operation of theoretical approaches on empirical inputs. They include: constant and systematic errors of technical measurement instruments (lens distortion in aerial cameras, atmospheric dust distortion, optical and electromagnetic measurements, temperature change altering the length of a physical measure); random and systematic (e.g. spatial autocorrelation) errors in statistical analysis; deficiencies in specification or calibration of mathematical models (in terms of overall fit, and in terms of specific refinements). In recognition of such factors, criteria of precision and accuracy are now routinely scrutinised in a number of fields. Their inclusion in the current framework is a means of scrutinising the correctness (appropriateness) of the precision represented.

Accuracy seeks to gauge whether the science has achieved a 'true' representation of the real-world phenomena under consideration. In some cases conventional goodness of fit statistics are (or can be) built into quantification processes. A 99% confidence limit would be 'good'; 95% might be 'fair', and so on, according to context. In other cases, however, the question of accuracy cannot be answered conclusively, or even directly, either because of inability to 'observe' the reality directly (for example, in forecasting contexts), or because of lack of agreement about suitable terms in which comparisons with reality should be made. Such difficulties are better acknowledged than ignored. It is also worth noting the trade-off: a quantitative estimate given originally as a range may warrant a higher 'accuracy' rating than one given as a point estimate, or a narrower range, for the former has more scope for spanning the 'true' value.

The finer the scale of measurement, the greater degree of precision being represented (parts per billion compared to parts per million; seven versus two significant digits). From a quality assurance point of view, it is necessary to know that the scale of measurement is appropriate for the phenomenon being represented. Rogue examples include the publication of indicators to five or six significant digits when many of the source statistics were more coarsely specified, or reporting of chemical pollutants to a scale that is beyond their limits of detection. It is also necessary to know that rounding errors are valid and whether point estimates have been given when ranges or intervals would have been more appropriate.

Errors within the margins of distortion already allowed for in the degree of precision adopted for representing the result need no further consideration. For

those that are not, it should be a matter of normal practice to incorporate appropriate correction factors, or specify error bars, confidence margins or other conventions in order to make due acknowledgment of them (these are automatically given in many statistical techniques, though not always rigorously implemented). Where this is done, a high precision score will be achieved. Where it is not, the score will be correspondingly low. Where precision is inherently problematic, qualitative representation of scientific outputs may be better than quantitative (numerical) expression of findings.

9.5.4 Peer review

This aspect captures one of the basic elements of the development of scientific knowledge – that of peer acceptance of the result. It is not sufficient for an individual or private agency simply to perform scientific investigations within their own terms and without a broader view. To claim a contribution to scientific knowledge, the result must be accepted across a peer community of appropriate independence and standing. The truth claim of any knowledge can only ultimately be assessed via discourse, and ultimately through consensus. Peer review is a fundamental element of the development of scientific knowledge.

In practice, review may be limited to self-appraisal, or a private group (as with consultancies and industrially funded and commercially confidential work), or it may extend quite widely to independent verification within a full, international peer community. It is also necessary to know about the outcome. The result may achieve widespread acclaim and endorsement. On the other hand, it may be severely criticised and even ridiculed.

The second of the theoretical aspects (state of the art) operates at a deeper level than the first (theoretical base) and provides a contextual backcloth for the latter. It sets out what can be expected in the light of the state-of-the art of a given field. One cannot expect to find well-tested theories in an embryonic field, and may need some convincing argument to tolerate mere speculation from an advanced field. The range is given from mature to ad hoc.

9.5.5 Validity

This invites assessment of the appropriateness of an estimate to the ‘real world’ problem to which it ostensibly relates, i.e. policy relevance. As is widely appreciated, model resolutions can be frustratingly deficient; models valid only for short-term projections are called on to produce long-term scenarios; highly aggregated generalised models are used for specific inferences; serious

mismatches can arise between the questions that risk managers need to address and issues that science can articulate. In some cases there may be ambiguity and a lack of consensus over the appropriate measure or indicator for a given problem. Owing to an absence of definitive context-specific knowledge about particular instances of environmental risk, it is often necessary to draw on knowledge from contexts believed to be similar in deriving risk estimates.

Experiments on animals under laboratory conditions may be the best available source of knowledge about the effects of certain radioactive isotopes on human beings (to conduct corresponding experiments on humans would be forbidden on ethical grounds). However, what remains unknown is the degree of transferability of that knowledge to humans under non-laboratory conditions (or even to the same species and type of animal under non-laboratory conditions). To take a different kind of example, historical data may be the best available source of information about certain sorts of failure rates of buildings, but again the degree of transferability of that knowledge to present-day conditions is unknown. And by way of further example, simulation models are by definition an artificial representation of a phenomenon or system of interest. A trade-off here with other aspects is very evident. The requirement for policy relevance can place unachievable demands on data quality.

If the logic tree used to represent the possible pathways of risk is incomplete (i.e. possible cause-effect links are missing) then this will critically undermine the assessment of risk. Many hazard incidents have occurred because of such omissions i.e. unforeseen possibilities. For example, the Exxon Valdez oil tanker crossed over a buffer lane, a lane reserved for incoming tankers, and an additional stretch of open water, before coming to grief. These had not previously been identified as credible events. At the Three Mile Island nuclear power plant, the valve failed to close (though an instrument panel showed that it had closed); again this had not been previously identified as credible. The Cleveland industrial fire in 1944 caused 128 deaths because the consequence of a spill with no containment had not been foreseen and therefore had not been built into the risk estimates.

Circumstances that render risk assessment particularly vulnerable to ‘completeness’ pitfalls (Freudenberg 1992) are:

- When the system is complex
- When there are gaps in knowledge about low probability events
- When there are substantial human factors
- When the system is untestable – an inherent characteristic of the real world settings of many waterborne risk contexts.

9.5.6 Summary

The 5 aspects have generated a total of 11 criteria against which the quality of risk assessment science can be examined. Risk estimates can be evaluated with respect to each of these criteria, generating a string of 11 scores. High scores will be a cause for comfort as they indicate a strong mature science, of direct policy relevance. Low scores will be a cause for caution, as they indicate science that has acknowledged weaknesses. Although a cause for concern and caution, they should not be a cause for shame or concealment – they are simply a measure of where we are – it is not necessarily possible to do any better.

9.6 REPRESENTING THE OUTPUTS

The simplest form of representation of the outcome of applying the above framework is as a string of scores. These, in turn, might be depicted graphically by way of a more immediately accessible visual representation. Figure 9.3, for example, is a graphical representation of the results from applying the current quality audit framework to two different sets of values for drinking-water consumption. Roseberry and Burmaster (1992) report a well founded sampling method, present upper and lower bounds for monitored consumption levels, and their results are now widely quoted and accepted. The US Environmental Protection Agency (EPA) figure of two litres has filtered into relatively widespread use, although its provenance is not a matter of verifiable record. Note that Figure 9.3 (along with Figure 9.4) has a total of 12 criteria, because it was based upon an earlier version of the framework, before ‘extent’ and ‘acceptance’ were combined to form a single category.

If a single, aggregate, indicator is required, the scores from each criterion can be added together, converted into a percentage rating, and evaluated against some standard set of benchmarks, to represent the degree of comfort that can sensibly be placed in the result, for example:

0–20%	Poor
20–40%	Weak
40–60%	Moderate
60–80%	Good
80–100%	Excellent

An aggregate score of 28 out of a possible 44 would translate to 63.5%, and its strength could accordingly be reported as being ‘good’. In the case of the results given in Figure 9.3, the aggregate score for the Roseberry and Burmaster

study is 37.5 out of a possible 48 (based on 12 criteria) yielding a rating of 78% (good); for the US EPA data, on the other hand, the aggregate score is 10.5, yielding a rating of 22% (weak).

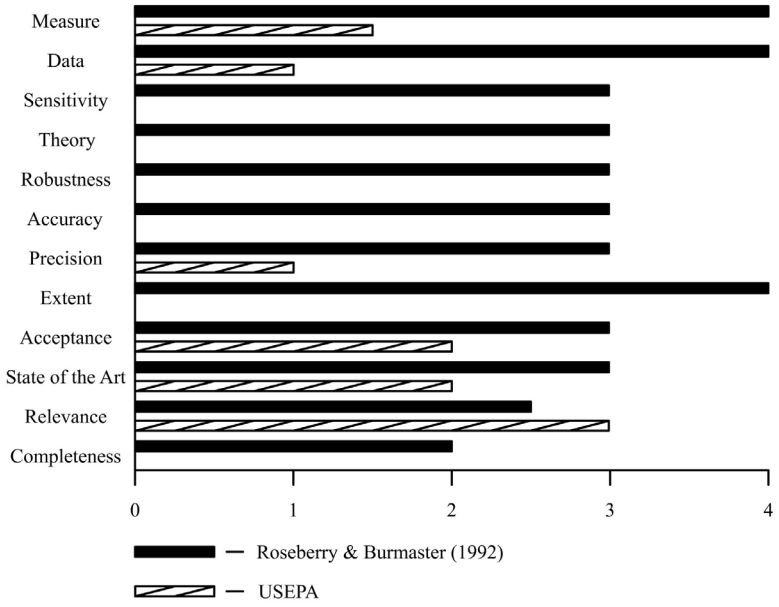

Figure 9.3. Outline quality audit of two different studies on drinking water consumption.

Not all criteria will necessarily be applicable in every context. Moreover, if they are all applicable, it may be appropriate to give a different relative weighting to each in the aggregation (as with weighted average multi-criteria methods more generally).

Where different types of scientific input are used in combination in a risk assessment, the issue arises of whether the quality audit should be applied to the composite result for the system as a whole, or whether distinct quality audits should be undertaken for individual components of the system in turn. In the former case (a composite audit) the audit process itself may be kept to manageable proportion overall, but the nature of the constituents may be so mixed as to make it difficult to apply the criteria in a meaningful way. In the latter case (a series of audits on individual components) the audit process will need to be repeated several times, but each application should have a coherent focus. The question of how best to combine the outputs of multiple audits raises further issues. For now, it is suggested that a 'weak link' rule is appropriate, in other words, the lowest score is taken for each category and the final assessment

is based on a table composed of such scores (see Macgill *et al*. 2000 for an example).

9.7 APPLICATIONS

The authors have applied the framework (or variants thereof) to a range of different examples of the assessment of waterborne risks. A high degree of convergence between different experts as to the criterion scores for specific cases has been found.

Its application to the determination of C*ryptosporidium* risks in drinking water demonstrated stark differences in the strength of available knowledge at three different points along the pathways through which human health risks may be generated (Fewtrell *et al*. 2001). Notably, it is considerably weaker at the consumer's tap (the point of exposure to risk) than at the treatment works, or in terms of environmental monitoring of raw water sources. These findings are summarised in Figure 9.4 (note 12 criteria rather than 11).

9.7.1 Quality audit case study

To illustrate a quality audit in a full format, rather than the summarised results, an example is taken from the wastewater reuse field. A summary of the study is presented followed by an outline audit, showing the reasons behind individual scores.

A study of the health effects of different irrigation types (raw wastewater, reservoir-stored water and rainwater) in agricultural workers and their families was undertaken in Mexico (Cifuentes 1998). The health outcomes examined were diarrhoea and infection with Ascaris. The case control study examined a total of 9435 people over a five-month period. In addition to collecting health and water quality data, information on potential confounding factors (such as socio-economic status, water supply, sanitation provision and so on) was also collected. The raw wastewater and rainfall irrigation areas were well matched in terms of housing conditions, mother's education, water storage and toilet facilities. The principal differences between these groups were the greater proportion of landless labourers in the raw wastewater group and the greater proportion of cereals grown in the rainfall area.

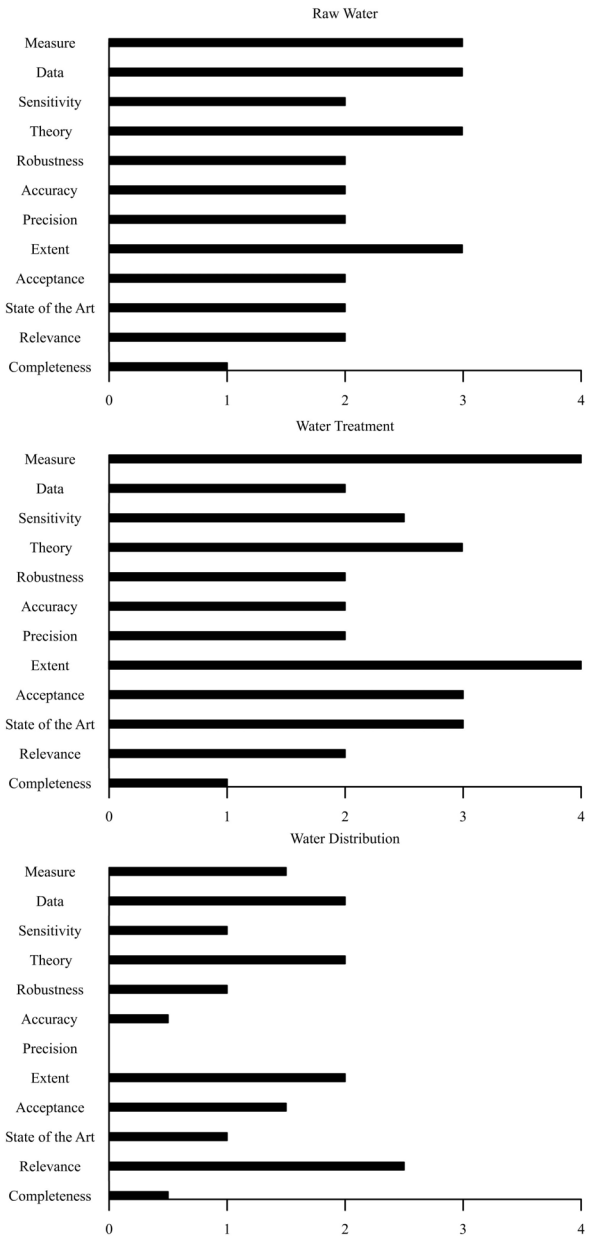

Figure 9.4. Outline comparative quality audit of three stages in the pathway of water supply.

Table 9.3 shows the outline quality audit for this study, based on the ascariasis outcome and the use of the study in terms of feeding into the guidelines process (in terms of 'Validity').

Table 9.3. Outline quality audit of wastewater reuse and levels of ascariasis

	Comments/level	Score
OBSERVATION		
Measure	Cases of ascariasis are being determined through faecal sample examination, and compared according to irrigation type.	
	Primary	4
Data	The empirical content is high, with power calculations conducted prior to the study to establish a suitable sample size.	
	Direct/Good	3
Sensitivity	Taking more than one sample per person may have increased the chances of finding positive cases. Other confounders, not accounted for, may be important.	
	Variable	2
METHOD		
Theory	The idea that pathogens can be isolated from faecal samples is well established, as is the idea that such pathogens may be transmitted via water.	
	Well-tested theory	3
Robustness	This is likely to be reasonable.	
	Variable – Resilient	2–3
OUTPUT		
Accuracy	This is certainly plausible if not better, with account taken for a number of known confounding factors.	
	Plausible	2
Precision	This is appropriate.	
	Fair	2
PEER REVIEW		
Extent	This type of cross-sectional study has been reviewed and, with appropriate note of confounding factors made, is fairly well accepted.	3
State of the Art	**Good**	3
VALIDITY		
Relevance	In terms of guidelines this study is directly relevant.	
	Direct	4
Completeness	The study examined a complete population, accounting for a number of confounding factors. It does, however, only relate to a small geographical area.	
	Partial – Majority	2–3
TOTAL		**31**

The quality audit result of 31 out of a possible 44 (i.e. 70%) demonstrates that it is considered that the study is well conducted, appropriate and can be used with a high degree of confidence. The reasoning behind each individual score is clearly laid out and can be used to stimulate discussion.

9.8 CONCLUSIONS

The case for quality audit of science for environmental policy is increasingly strong. It is not sufficient for experts intuitively to appreciate various areas of uncertainty in terms of which their findings should be qualified. Accountability calls for the evidence to be formally represented, so that all stakeholders can formulate a responsible view. Robust tools are needed for the job. In developing and testing such tools, there will inevitably be a need for compromise over the ideals of simplicity and transparency, on the one hand, and that of achieving a faithful representation of the complexities and subtleties of scientific endeavour, on the other. The framework presented here is offered as a practicable solution that can be the basis for further development and refinement in the future. Such development may include its formulation within interactive communication and information technology systems, in order to facilitate access and deliberative participation on the part of a wider group of experts in arriving at appropriate criterion scores for particular cases.

In summary, the framework outlined here allows outcomes of the risk assessment procedure to be a more transparent process open to scrutiny. Individual quality audit tables also highlight areas that could be improved and provide a platform for debate. Following the QA framework procedure through the risk assessment process should also allow decisions to be updated more easily, since only areas where there have been significant changes need to be re-examined and the results combined with the original assessment.

Widespread adoption of the QA process should prevent numbers from developing a life of their own. It is the antithesis of science to hide data imperfections and doubtful assumptions; on the contrary, there should be openness. There should be no shame in saying ‘it’s the best there is at the moment’ (if of course it really is the best and not just something being used for convenience). If nothing else, then the foundation for the eternal plea for ‘more research’ will have been clearly established.

9.9 IMPLICATIONS FOR INTERNATIONAL GUIDELINES AND NATIONAL REGULATIONS

International guidelines provide a common (worldwide) scientific underpinning; as such, it is increasingly necessary to have a rigorous quality control procedure. At present, reliance is placed on the quality implied through the peer review process. The idea of a predefined and systematic quality review such as the one defined in this chapter essentially levels the playing field and allows judgements to be made from a common starting point. Such a systematic framework is also valuable at national levels as it provides a means by which unpublished data can be evaluated. Development (by the WHO) of a complementary framework or scoring system outlining the overall strength of evidence and coherence of inputs to international guidelines is underway. Together these will provide valuable input to guidelines and standards development and will also aid in the risk communication process.

9.10 REFERENCES

Burmaster, D.E. and Anderson, P.D. (1994) Principles of good practice for the use of Monte Carlo techniques in human health and ecological risk assessments. *Risk Analysis* **14**(4), 477–481.

Cifuentes, E. (1998) The epidemiology of enteric infections in agricultural communities exposed to wastewater irrigation: perspectives for risk control. *International Journal of Environmental Health Research* **8**, 203-213.

Cranor, C.F. (1995) The social benefits of expedited risk assessment. *Journal of Risk Analysis* **15**(3) 353–358.

Fewtrell, L., Macgill, S., Kay, D. and Casemore, D. (2001) Uncertainties in risk assessment for the determination of drinking water pollutant concentrations: *Cryptosporidium* case study. *Water Research* **35**(2), 441–447.

Freudenberg, W.R. (1992) Heuristics, biases and the not so general publics. In *Social Theories of Risk* (eds S. Krimsky and D. Golding), pp. 229–249, Praeger, Westport, CT.

Funtowicz, S.O. and Ravetz, J.R. (1990) *Uncertainty and Quality in Science for Policy*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Gale, P. (1998) Development of a risk assessment model for *Cryptosporidium* in drinking water. In *Drinking Water Research 2000,* Drinking Water Inspectorate, London.

Haas, C. and Eisenberg, J. (2001) Risk assessment. In *Water Quality: Guidelines, Standards and Health. Assessment of risk and risk management for water-related infectious disease* (eds L. Fewtrell and J. Bartram), IWA Publishing, London.

Haas, C.N. and Rose, J.B. (1994) Reconciliation of microbial risk models and outbreak epidemiology: The case of the Milwaukee outbreak. *Proceedings of the American Water Works Association Annual Conference, New York,* pp. 517–523.

Havelaar, A.H. (1998) Emerging microbiological concerns in drinking water. In *Drinking Water Research 2000,* Drinking Water Inspectorate, London.

Macgill, S.M., Fewtrell, L. and Kay, D. (2000) Towards quality assurance of assessed waterborne risks. *Water Research* **34**(3), 1050–1056.

Macler, B.A. and Regli, S. (1993) Use of microbial risk assessment in setting US drinking water standards. *International Journal of Food Microbiology* **18**, 245–256.

Medema, G.J., Teunis, P.F.M., Gornik, V., Havelaar, A.H. and Exner, M. (1995) Estimation of the *Cryptosporidium* infection risk via drinking water. In *Protozoan Parasites and Water* (eds W.B. Betts, D. Casemore, C. Fricker, H. Smith and J. Watkins), pp.53–56, Royal Society of Chemistry, Cambridge.

NAS (1983) *Risk Assessment in the Federal Government: Managing the Process*, National Academy Press, Washington DC.

Perz, J.F., Ennever, F.K. and le Blancq, S.M. (1998) *Cryptosporidium* in tap water. Comparison of predicted risks with observed levels of disease. *American Journal of Epidemiology* **147**(3), 289–301.

Rose, J.B., Lisle, J.T. and Haas, C.N. (1995) Risk assessment methods for *Cryptosporidium* and *Giardia* in contaminated water. In *Protozoan Parasites and Water* (eds W.B. Betts, D. Casemore, C. Fricker, H. Smith and J. Watkins), pp. 238–242, Royal Society of Chemistry, Cambridge.

Roseberry, A.M. and Burmaster, D.E. (1992) Log-normal distributions for water intake by children and adults. *Risk Analysis* **12**(1), 99–104.

Weinberg, A. (1972) Science and trans-science. *Minerva* **10**, 209–222.

Whittemore, A.S. (1983) Facts and values in risk analysis for environmental toxicants. *Risk Analysis* **3**(1), 23–33.