

mixed methods evaluation

The wheelbarrow, the mosaic and the double helix

Challenges and strategies for successfully carrying out mixed methods evaluation

Lois-ellin Datta¹

Introduction

What is the best way to learn whether a program phasing out direct income supports, while phasing in limited eligibility, employment transition, child care, health, housing and food benefits, works? The discussions at the July 2000 forum on impact evaluation held in Wellington, New Zealand, consistently indicated that there was no one best method to which one would turn first, the others being defaults. The golden rule, in Burt Perrin's wonderful phrase, is that there is no golden rule in evaluation methods.

Rather, the discussions emphasised that many approaches can be of great use, depending on factors such as the policy context, including key questions as seen by relevant stakeholders, program maturity, data availability and the desiderata to be maximised (conference summary prepared by the New Zealand Government, undated). It followed that scoping the evaluation could benefit from a multi-disciplinary team, from front-end listening to diverse stakeholder groups, and from early clarity on the evaluation desiderata to be maximized.

Some methods would shine against certain of the desiderata but have notable limitations for others. Randomised designs, such as Boruch (1997) presents, would be appropriate in some situations; quasi-experimental designs, such as those described by Lipsey and Cordray (2000), in others; theory-based and intermediate variable designs, such as Pawson and Tilley (1997) and Henry, Julnes and Mark (1998, 1999) discuss, in still others. As a reminder, Table 1 illustrates some of these desiderata.

Many evaluations involve multiplism. The discussions also emphasised that in many instances mixed methods would be required. This point is underscored in a forthcoming discussion on evaluations of welfare reform by Julnes and Foster (2001). They argue for using multiple methods to provide a more comprehensive view of the outcomes experienced by welfare leavers, citing Cook's (1985) explication of critical multiplism. As seen by Cook, critical multiplism could involve:

- multiple value stances;
- multiple program theories;
- multiple operationalisation of constructs;
- multiple methodology paradigms;
- multiple professional affiliations of investigators;
- multiple contexts for inquiry.

This is hardly a tough argument to make. Rather, it seems fair to say that much current evaluation practice involves, routinely and de minimis, multiple types of data.

Lois-ellin Datta is a past President of the Evaluation Research Society (now the American Evaluation Association), and is President of Datta Analysis.

Often, for example, an evaluation will use time-series analyses of outcome monitoring or administrative performance data with interviews or surveys, together with case studies including observations of participants. Frequently, an outcome or impact evaluation in addition will involve multiple methodologies such as ethnographic or economic perspectives, multiple professional affiliations of investigators, and multiple contexts such as sites with quite divergent characteristics or cross-national comparisons for inquiry. Such combinations seem quite accepted for national evaluations by policy analysts and decision-makers.

Despite the considerable emphasis on the rationale for mixed methods and the wide spread acceptance of an array of methods, there is relatively little systematic information available on how, successfully, to carry out a mixed methods evaluation, from management through data collection to analysis and reporting.

Obviously, evaluators have developed heuristics, ways learned from experience on how to conduct mixed methods studies and deliver reports on time. One would expect that such approaches will be modified, as craft knowledge develops and is shared.

However, what one might think are some basic questions are just beginning to be more systematically addressed. Further, based on reading

Despite the considerable emphasis on the rationale for mixed methods and the wide spread acceptance of an array of methods, there is relatively little systematic information available on how, successfully, to carry out a mixed methods evaluation, from management through data collection to analysis and reporting.

a fair number of completed evaluations, there may be room for clarification, and possibly improvement, on whether a wheelbarrow, a mosaic, or a double-helix is needed and achieved.

Learning from experience

Some recent papers and a symposium held by the MacArthur Foundation in January 2000 are bringing together experience with the pitfalls and possibilities in the practical aspects of using mixed methods.

Assume that proper consideration has been given to such factors as the availability of evaluators skilled in the methods of choice, the greater costs usually involved in data collection, the greater time usually demanded for analysis and interpretation, and how different methods may be more or less brittle should key staff leave. Assume too that the advantages – and there are many – of mono-methods have been thoroughly, fairly considered, but the decision has been made that answering the evaluation questions requires mixed methods. Now the ship must be built, launched, and guided successfully on the voyage.

Helpful definitions

Two definitions may be helpful in discussing these issues.

The types of analysis

First, I'll refer to two types of analysis: cross-over tracks and parallel tracks. These are illustrated in Figures 1 and 2.

In *cross-over tracks analysis*, findings from the various methodological strands intertwine and inform each other throughout the study. In *parallel tracks analyses*, the analyses are conducted independently, according to standards of quality and excellence for each method. The findings are brought together after each strand has been taken to the point of reaching conclusions (Li, Marquart & Zercher 2000).

Wheelbarrow, mosaic and double-helix

Second, I'll speak of the 'wheelbarrow,' the 'mosaic' and the 'double-helix.'

- The *wheelbarrow* is method-driven, often presenting the group commissioning the evaluation with separate reports such as richly textured individual site findings, possibly with a summary report of major themes across sites; reports from various country-wide data sets; a summary report of major findings across these national data sets; perhaps a report summarising the summaries.
- The *mosaic* is also method-driven in keeping the findings from various approaches fairly distinct. However, the final report brings together as much as possible of the findings from these approaches, perhaps using the more qualitative or case study data as vignettes, illustrations and ways of putting a human face on the national data sets; perhaps as thematic findings.
- By *double-helix*, I mean a question-driven report. Here methodology is backstage, in an appendix, and the report is organized by key questions and their answers, integrated from the beginning across methods: 'What was the operational logic? Was the program carried out as intended? When it was carried out, what were the results? Those intended? Other unintended changes? For which subgroups if any was the program differentially effective? Can the results reasonably be attributed to the program?'

FIGURE 1: ANALYTICAL FRAMEWORK FOR THE PARALLEL TRACKS ANALYSIS

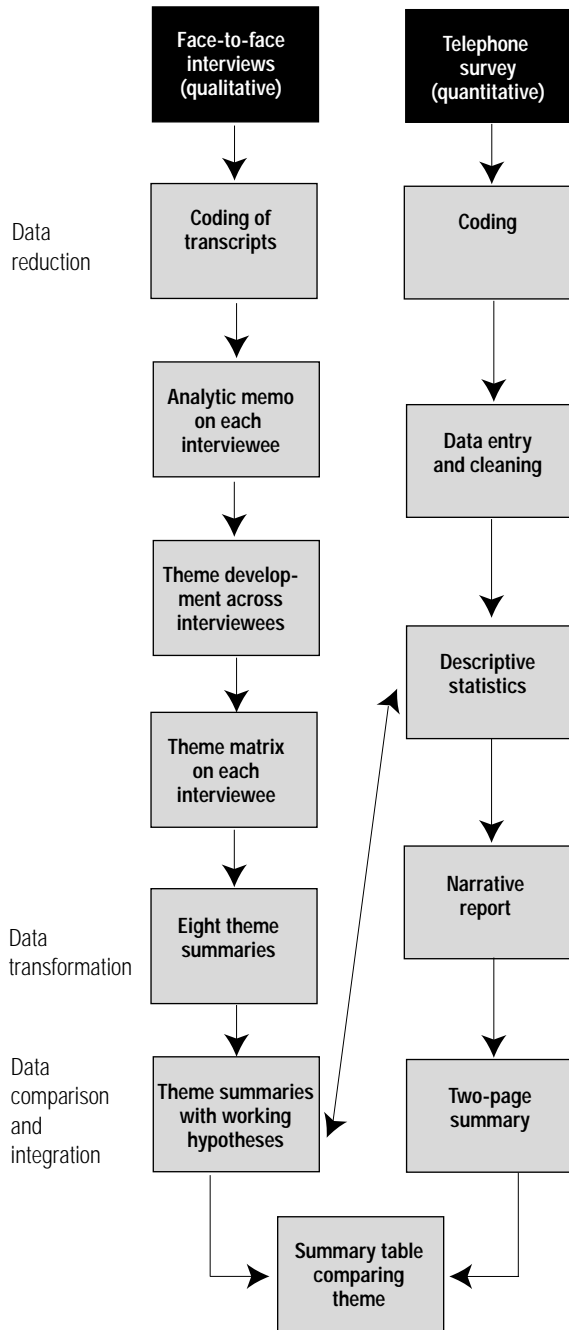
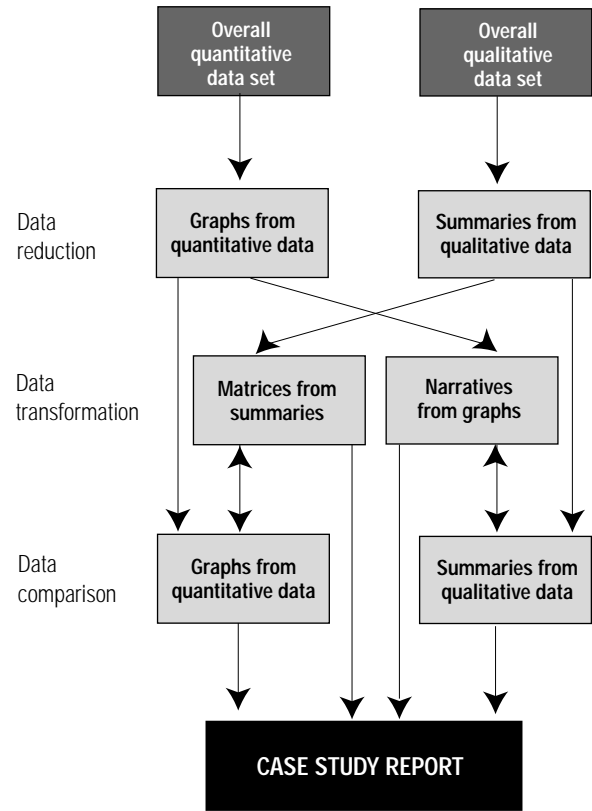


FIGURE 2: ANALYTICAL FRAMEWORK FOR THE CROSS-OVER TRACKS ANALYSIS



Source Figures 1 and 2:
 Li, S., Marquart, J. M & Zercher, C. 2000, 'Conceptual issues and analytical strategies in mixed-method studies of preschool inclusions', *Journal of Early Intervention*, no. 20, pp.116-132

helpful definitions

The next sections look at the challenges – which can be ferocious – and about strategies for successfully carrying out mixed methods studies.

Here be tigers

Challenges and pitfalls in carrying out mixed-methods studies

Perhaps one of the profiles in courage in our field was the decision by an evaluation firm, Abt Associates, Inc., to give a senior analyst paid leave to write a paper about a cataclysmic disaster in mixed-methods use. This paper (Trent 1978), begins

The new question was how a program could produce such splendid results in so many aspects when the observation data consistently indicated the program would be a failure. This analysis was done by the top leaders, the senior analysts, of both teams.

with the story of three social experiments supported by the US Department of Housing and Urban Development to see if using a direct cash housing allowance would help low-income families find decent housing on the open market. It was hoped the full-scale allowance program would expand the local housing market in urban areas and improve freedom of choice by allowance recipients, leading to greater efficiencies in housing stock use. Prior to and during the experiments, the status quo was twofold: support for building large new housing blocks for low-income families, and support to landlords who rented to eligible families. The paradigm shift being tried out was placing greater control in the hands of the families themselves.

The evaluation design was quasi-experimental: comparison of housing outcomes in the experimental sites to those of matched 'status quo' sites. Measures included quantitative and qualitative data. Six common sets of forms organised administrative data and tracked the progress of participating families. Cost-effectiveness was a salient evaluation question, so local employees such as housing counsellors and housing inspectors noted their charges for time. Agency accountants tracked their charges. There were, in addition, interviews with samples of participants and those leaving the program, before and after, to see if those receiving benefits moved to better housing. On the qualitative side, anthropologists were assigned to each site to watch program operations on a daily basis, including agency staff and participants.

The analyses were conducted as parallel tracks: the quantitative data by one team and the qualitative data by another team, with the findings to be integrated in a single final report: the aim was a double-helix. The details are given in admirable clarity in Trent's article ... right up to the point where the two locomotives crashed into each other.

In essence, the quantitative data indicated program success, including particular success at one important location, Site B, designated as the

'flagship' project for the initiative. Here, however, the qualitative observer concluded that the site B was fraught with conflict, was an operational disaster as seen by the observer, involved de-facto racism, and could not plausibly have led to the good results indicated by the quantitative analyses.

The evaluators could have swept this under the rug by the fairly common strategy of simply reporting the two findings as different perspectives after some checks to see if there were some reasonably apparent explanations for the differences – a mosaic report. To their credit, the teams struggled to obtain fair, independent reviews of both qualitative and quantitative data on site B.

The site B observer was asked to reorganise his paper, which was described by the reviewers as 'engaging but focussed on only personality conflicts and managerial competence to explain what was going on at the site.' Five months later, the problem still remained: the observer continued to conclude the program was disastrous, but the outcome computer runs showed the site had the second-best outcomes with regard to cost-effectiveness in the whole study. The Abt evaluation staff now was polarised, with the qualitative team defending their colleague with statements such as '... the outcome measures did not tell the whole truth, quantitative techniques were garbage and human behaviour could not be reduced to mere numbers'. The quantitative staff returned the compliment with interest with regard to qualitative data and methods.

So the task was recast: instead of seeing what was wrong with two approaches, the question was restated to assume both methods had been properly carried out and were methodologically first-rate. The new question was how a program could produce such splendid results in so many aspects when the observation data consistently indicated the program would be a failure. This analysis was done by the top leaders, the senior analysts, of both teams.

Eight weeks later, by breaking down Site B into sub-sites and a more fined-grained analysis using the cross-over tracks techniques, some answers emerged: the horrific situations that the observer accurately noted had worked in the agency's favour. These included staff leaving, which meant their salaries could be used to cover other expenses as vacancies were not filled. Remaining staff carried dual workloads. These 'mass-production' methods gave an illusion of efficiency at the cost of a high dropout rate among Black enrollees. Some hunches could not be adequately tested by the data. The 'essence' paper for Site B that finally emerged had no real heroes or villains and was used to re-examine other sites. Eventually – almost a year late – the effort yielded a report the company was ready to stand behind. As Trent dryly notes, 'Staff began to think of ways that conflicting interpretations could be used for gains in understanding'.

The challenges in making mixed methods work

successfully may be most apparent at the data interpretation stages, but are by no means limited to these. At the centre are tensions around allocation of resources and varying degrees of trust in less familiar methods. That these are still with us is illustrated by Goldenberg, Gallimore and Reese's (2001) conclusion at the end of their mixed-methods exploration of intergenerational influences on adolescent development:

This story, with its ambiguous two-option ending, illustrates the outcome of a choice we made at the inception of this research program. Our methodological ecumenism led us to seek out colleagues and fieldworkers with diverse empirical perspectives and we encouraged the intellectual diversity this produced. This choice both benefited our research and led to tensions that endangered portions of it...Nonetheless, we remain committed to the belief that [this set of questions] require multiple and complementary methodologies.

And here be springs of sweet water

Some approaches to avoiding the pitfalls

In January 2001, the MacArthur Foundation brought together researchers and evaluators who had used mixed methods successfully in studying

pathways through middle childhood. The purpose was to share what is known about how to make this approach work, by examining exemplary evaluations in some depth. The study I selected was conducted by Mary Ann Millsap (qualitative) and Ann Chase (High Church quantitative) of Abt Associates, examining a large, longitudinal test of an approach by James Comer to school reform (Millsap 2000). It was exemplary in terms of its findings as an instance where mixed methods averted inappropriate death by control groups, and it was exemplary in its management approaches to making mixed methods work.

Tables 1A, B and C summarize strategies identified at the conference and elsewhere (Bamberger 2000; Brock 2001; Gibson & Duncan 2001; Datta 2001, Greene and Caracelli 1997; Kidder & Fine 1987). Most centre on study management; some involve analytic techniques.

In the area of management (Table 1A below), the underlying theme is developing a deeper understanding than is usual of each of the methods involved, hopefully leading to a deeper than usual appreciation of the strengths and limitations of each. For example, in the study of the Comer program, each of the in-depth, longitudinal site visit teams included a member of the team primarily responsible for the administrative data, survey and test analyses. These quantitative team members participated fully in being trained for the site visits, writing up reports, and analysing the site visit data.

TABLE 1A: STRATEGIES FOUND HELPFUL IN MAKING MIXED METHODS WORK
– management and fostering cross-team confidence in the data

Strategy	Details
<i>Co-project managers</i>	The evaluation direction could be vested equally in evaluators from different methodological heritages skilled in the key methods used.
<i>Staffing</i>	Staffing could be approximately equal in number, experience, and status for the key methods used.
<i>Decision-making protocols</i>	Protocols could be developed before the evaluation moves to implementation for ways in which the teams will invest time to work together before decisions are made and for decision-making.
<i>Cross-team assignments</i>	A senior analyst from one team could serve also on the other team(s) and vice versa. When there is first-hand knowledge of how data are collected, then there is greater respect for the multiple approaches.
<i>Cross-track cooperation and cooperative activities</i>	Questions addressed in site visits (for example) can be elaborated and expanded on in surveys; survey and other database findings can help direct observations, interviews and other types of approaches; work on development of measures such as survey forms or observation schedules is done by a mixed-method team.
<i>Building a common vision</i>	Front-end time in developing a common vision among team members about the evaluation purpose and context; promoting frequent communications among staff at all levels and from the key methodological traditions, through approaches such as e-mail distributions, briefings and seminars, cross-training.

TABLE 1B: STRATEGIES FOUND HELPFUL IN MAKING METHODS WORK
– technical protocols affecting analysis

Strategy	Details
<i>Early decisions on weighting</i>	Early in implementation, decisions could be made on the relative weights to be given to each data element in reaching conclusions. If possible a fairly formal procedure should be used, such as Scriven presents.
<i>Early conceptual clarity on whether model is triangulation or a mosaic</i>	Triangulation makes sense when two or more data sources are considered independent takes on the same question, intended to cancel out biases. A mosaic image makes sense when two or more data sources are considered as addressing different issues or questions.
<i>Early decision on inferences and actions when data intended as triangulation do and do not agree</i>	Both agreement and disagreement in a triangulation situation can be misleading. The agreement, when looked at carefully, may be due to shared errors and would not justify greater strength of conclusions. The disagreement may be due to insufficiencies in data analysis, particularly analytic grain. Early on, decisions should be made on steps to be taken in both of these situations.
<i>Parallel track analysis</i>	Particularly in triangulation situations, where convergence would be expected. Promotes development of different interpretations, avoids premature closure, and helps assure proportionate weighting of findings from different approaches.
<i>Cross-track analysis</i>	Particularly in mosaic situations: expands scope of inquiry, enriches understanding of the event, shows the 'whole picture' as answers to different questions develop.

TABLE 1C: STRATEGIES FOUND HELPFUL IN MAKING MIXED METHODS WORK
– reporting where a double-helix is sought

Strategy	Details
<i>Writing</i>	Co-mingling report preparation. For example, the primarily quantitative sections may be written by the primarily qualitative co-director and team and vice versa. The structure of the report focusses on answers to the evaluative questions and is not organized by the methodology. The methodology appears in an appendix, rather than being prominent in the report text. 'Ownership' of the findings, interpretation, conclusions, recommendations is independent of discipline or method.
<i>Presentation</i>	Each of the co-directors can report fully on all aspects of the evaluation. Presentations may be organized by answers to the evaluative questions. They are not segmented by methodology: findings from the outcome, performance analyses by one co-director and findings from the interview or observational data by the other co-directors.

(Source Table 1A, 1B, 1C: Li, Marquart and Zercher, 2000)

And the site visit teams participated fully in collecting, cleaning and analysing the large databases. Mathophobia was not permitted!

In the area of technical protocols affecting the analysis (Table 1B), the underlying approaches involve early clarity on issues such as weighting, whether the model is one of data and method triangulation, or one of data complementarity where convergence would not be expected, and particularly whether the overall approach would be parallel track or cross-track analysis. That is, in some situations, such as interviewing labour union stewards, employees and managers in a study of employee relations, one does not expect views to converge within a method. In other situations, convergence among these perspectives would be expected and sought. Likewise, in some situations, convergence in findings across methods such as time-series analyses of monitoring data and case studies would not be expected; in others, they would. To make mixed methods work well, it can be useful indeed to be crystalline at the beginning of a study about whether the multiplism is intended as a mosaic – different, not necessarily convergent views – or as triangulation, i.e. increasing certainty through

methods that cancel out each other's limitations.

We are reminded that even in the usual scrunch of getting out the final report, one looks across findings and reports to identify

convergences, complementarities and divergences. In most circumstances, one does not want to dump the wheelbarrow on the policy analyst to sort out the sense. However, we also are reminded that one needs to shake the findings tree thoroughly to be as sure as possible that convergence in a triangulation situation is not the result of shared error or common biases – and to see whether apparent divergences reflect unexpected clusters or subgroups that vary across sites or across data sets.

Both notions glide down the cortex like jelly: easy to say. They can turn out to be in practice quite difficult to do for multi-site, multi-method evaluations because data collection, analysis and report writing seem often to take twice the time allowed, and the reporting dates can slip only so much.

With regard to reporting, the comments about time and about front-end clarity as to whether the design, conceptually, is mosaic or double-helix – separate perspectives or triangulation – become

To make mixed methods work well, it can be useful indeed to be crystalline at the beginning of a study about whether the multiplism is intended as a mosaic – different, not necessarily convergent views – or as triangulation, i.e. increasing certainty through methods that cancel out each other's limitations.

underscored. The wheelbarrow is perhaps the easiest to prepare, and, as noted, can be necessary and invaluable for responsible communication and consultation with individual sites. The double-helix is perhaps most natural for those trained as policy analysts and more difficult for those trained as economists, sociologists or evaluators. The strategies suggested, such as having the team leaders for different sectors actually write each other's 'sections', are not particularly efficient, at least at first. The results, however, particularly from the perspective of the public and the policy-maker, can be worth the effort.

In the Comer example, for instance, the initial analysis showed no value added by the program, despite five years of implementation effort and the persuasiveness of the program theory. The study was headed for another death by randomised control group report. The cross-team experience led, however, to more in-depth examination of degrees of implementation in both Comer and control sites, and to establishing that while the Comer ideas were becoming increasingly part of general school reform, value was added, quite substantially, for the schools when the concepts were thoroughly implemented.

This led, to be sure, to questions such as the practicality of the model if implementation proved difficult, relative to the motion that perhaps expectations should include model multiplicity.

Some day, all evaluators may be methodologically multilingual, knowing well and valuing diverse traditions, value stances, program theories, approaches to operationalising constructs and methodology paradigms. At least one evaluator approaches successful use of mixed methods by

selecting the data collection team leaders from among the rising generation of evaluators who have had such multilingual training. It is encouraging that some leaders now work hard at offering this training: qualitative evaluators such as Greene, quantitative such as Sechrest, and 'realistic' evaluators such as Mark. The strategies offered in this paper are a beginning. As more teams work together on the planning, analysis and reporting of mixed methods designs for national policy evaluations, many new approaches surely will be added, and our own certainty increased that our result will be – if this is appropriate for the policy and program at issue – a double helix.

Note

- 1 Since the 1960s when she became national director of Head Start evaluation, Lois-ellin Datta has been involved in planning, funding, carrying out, writing, teaching, fretting and rejoicing in evaluations. A pragmatist, her work has emphasised finding appropriate ways to answer the evaluation and policy

questions with maximum information for various stakeholders, minimum intrusion on the people and programs involved, and a deep respect for operational as well as program logic. A past President of the Evaluation Research Society (now the American Evaluation Association), she currently is President of Datta Analysis – and as of 15 November, a coffee farmer. Cuppa anyone?

References

- Bamberger, M. (ed.) 2000, *Integrating Quantitative and Qualitative Research in Development Projects*, Directions in Development Series, World Bank, Washington, DC.
- Boruch, R. 1997, *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Sage, Thousand Oaks, California.
- Brock, T. 2001, Viewing mixed methods through an implementation lens: A response to the New Hope and Moving to Opportunities evaluations, paper presented at the MacArthur Conference on Mixed Methods, January 2001.
- Cook, T. D. 1985: 'Postpositivist critical multiplism', in *Social Science and Social Policy*, eds I. & M. M. Mark, Sage, Thousand Oaks, California, pp. 21–62.
- Datta, L-e 2001, Avoiding death by evaluation: The Abt evaluation of the Comer Approach, paper presented at the MacArthur Conference on Mixed Methods, January
- Gibson, C. M. & G. J. Duncan 2001, Qualitative/quantitative synergies in a random-assignment program evaluation, paper presented at the MacArthur Conference on Mixed Methods, January 2001.
- Goldenberg, C., R. Gallimore & L. Reese 2001, Using Mixed Methods to Explore Latino Children's Literacy Development, paper presented at the MacArthur Conference on Mixed Methods, January 2001.
- Greene, J. C. & V. J. Caracelli (eds) 1997, *Advances in Mixed-method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms*, New Directions for Evaluation, no. 74, Jossey-Bass, San Francisco.
- Henry, G. T., G. Julnes & M. M. Mark (eds) 1998, *Realistic Evaluation: An Emerging Theory in Support of Practice*, New Directions for Evaluation, no. 79, Jossey-Bass, San Francisco.
- Julnes, G. & Foster, E. M. 2001, Crafting evaluation in support of welfare reform, unpublished paper.
- Kidder, L. H. & M. Fine 1987, 'Qualitative and quantitative methods: When stories converge', in *Multiple Methods in Program Evaluation*, eds M. M. Mark & R. L. Shotland, New Directions for Program Evaluation, no. 35, Jossey-Bass, San Francisco, pp. 57–75.
- Li, S., Marquart, J. M & Zercher, C. 2000, 'Conceptual issues and analytic strategies in mixed-method studies of preschool inclusions', *Journal of Early Intervention*, no. 20, pp. 116–132.
- Lipsey, M. W. & Cordray, D. S. 2000, 'Evaluation methods for social intervention', *Annual Review of Psychology*, no. 51, pp. 345–375.
- Mark, M. M., Henry, G. T. & Julnes, G. 1999, *Evaluation: An Integrated Framework for Understanding, Guiding and Improving Policies and Programs*, Jossey-Bass, San Francisco.
- Millsap, M. A. et al. 2000, *Evaluation of Detroit's Comer Schools and Families Initiative*, Abt Associates, Cambridge, Massachusetts.
- Pawson, R. & Tilley, N. 1997, *Realistic Evaluation*, Sage, Thousand Oaks, California.
- Trent, M. G., 1978, 'On the reconciliation of qualitative and quantitative analyses: A case study', *Human Organization*, vol. 37, no. 4, pp. 345–354.



Tunde Meikle, Genevieve Pépin and Helen Goodman (from left to right) won student scholarships to attend the 2001 Canberra International Conference.

Tunde and Helen were sponsored by the Centre for Program Evaluation, while Genevieve, who travelled from Québec, was assisted by the conference committee.