Pre-Publication Draft Subject To Revision

NC2 Methods Network Workshop

Systematic Reviews in Social Science

Some Essential Challenges

Finding a Workable Review Title: Intervention, Target Group & Outcome

Session 2: How to choose outcome measures

Opening Comments of Session Chairperson Edward J Mullen, Professor, Columbia University

Session Moderator: Larry V. Hedges, Professor, University of Chicago

Discussant: Edward Melhuish, Professor, University of London

November 10, 2004

Nordic Campbell Center

The Danish National Institute of Social Research

Copenhagen, Denmark

Recommended Readings for Workshop Participants

Alderson P, Green S, Higgins JPT, editors. *Cochrane Reviewers' Handbook 4.2.2* [updated December 2003]. Section 4, 8.1, 8.2 & 9. In: The Cochrane Library, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.

Sections 4 and 9 of the *Cochrane Reviewers' Handbook* provide guidance on how reviewers should plan for outcomes measurement in Cochrane reviews. Sections 8.1 and 8.2 look ahead to the analysis phase and provide an overview of various effect size measurement possibilities. While the content of section 8 will be addressed in a subsequent session it is useful to have these analytic possibilities in mind when formulating plans for outcomes measurement in the protocol.

Cooper, H., & Hedges, L. V. (Eds.). (1994). Parts I & II. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Part I provides a general overview of research synthesis. Part II addresses conceptual and statistical aspects of hypotheses and problem formulation in research syntheses. This is an important context for workshop participants to have.

Gray, J. A. M. (2001). Chapter 6: Assessing the outcomes found. In *Evidence-based Healthcare*. (2nd edition). Edinburgh: Churchill Livingstone.

This chapter is a guide for assessing outcomes and provides an easily understandable overview of some of the key concepts in the context of evidence-based healthcare. The book can be ordered on-line from Elsevier at: http://www.us.elsevierhealth.com/product.jsp?isbn=0443062889

The purpose of this workshop is to initiate an open and constructive discussion on the topic of developing a systematic review framework for the social sciences. Our day has been structured to address five challenges faced by social science researchers when planning and implementing systematic reviews. A first challenge is to formulate a workable review title. This includes three sub-challenges including how to delimit an intervention, how to specify the intervention target including problem and population, and finally how to choose outcome measures. This session is structured to facilitate an open and constructive discussion of the challenge of how to choose outcome measures. I am not going to comment on subsequent challenges, the challenges of gathering data about outcomes, analyzing that data, and interpreting the results. My assignment is to help frame the challenges so as to provide some context for the discussion. You may find my presentation a bit frustrating since I intend to identify challenges and to raise questions for subsequent discussion, but I do not attempt to provide answers.

Outcomes measurement defined

Before discussing challenges facing those planning systematic reviews I take a brief detour to define what I mean by *outcomes measurement*. First, I comment on the meaning of outcomes measurement in general and then with specific reference to systematic reviews.

The Cochrane Reviewers' Handbook 4.2.2 (Alderson P, Green S, Higgins JPT, 2004) states that a review question should specify not only the interventions and participants of interest, but also the types of outcomes of interest. The *Handbook* notes

that it is the contrast between the outcomes of two groups treated differently that is known as the *effect*. This is a subtle distinction but one that is useful to keep in mind in our discussions. The concepts of *effect* and *effect size* take us a step beyond what is meant by the term *outcome*. For instance, I can refer to increased employment as an outcome, but to invoke the term *effect* I must calculate some contrast such as between an experimental group and a control group. Ultimately, reviewers are interested in effects including effect direction, size, consistency and the strength of evidence for an effect.

In evaluation research wherein social program's are studied Rossi (1997) has described outcomes measurement as follows:

> The outcomes of a program or policy are changes, intended or not, in the program's targets that accompany exposure to the program. In human services programs, targets can be persons, families, neighborhoods, schools, agencies, firms, etc. to which the program is directed. Programs can cover a variety of activities designed to achieve intended outcomes, including the providing of information, counseling, material support, training, laws and legal sanctions, medical therapy, etc. (p. 21).

Donabedian (1981) defined health outcomes as changes in a patient's current and future health status that can be attributed to antecedent health care. This definition is widely accepted within healthcare. In the report *Australia's Health 2000*, health outcome is defined as "A health related change due to a preventive or clinical intervention or service. (The intervention may be single or multiple and the outcome may relate to a person, group or population or be partly or wholly due to the intervention)" (Australia

Institute of Health and Welfare, 2000, p. 444). The British National Health Service describes outcomes as “The attributable effect of an intervention or its lack on a previous health state” (United Kingdom Clearing House on Health Outcomes, March 1997).

Definitions of “outcomes” applicable to general public sector services are consistent with these health definitions. In the United States the Government Performance and Results Act of 1993 (1993) defines outcome as “…the results of a program activity compared to its intended purpose” (§1115). All of these references tie outcomes to identifiable, traceable interventions, at least in part.

In the context of systematic reviews outcomes measurement takes on a somewhat different character. The review question asks about outcomes associated with a specific intervention for a specific population/problem, aggregating evidence from multiple studies. The challenges in studying such outcomes differ from those faced by investigators conducting primary studies. My remaining comments outline what I see as some of those challenges.

Challenges for authors of systematic reviews

I have been asked to identify challenges for this session’s discussion. I am sure that many challenges will be mentioned before we conclude this morning. I have selected eight that I consider most relevant. These have to do with:

1. Which outcomes constructs to include and exclude in the review question
2. Which measures of outcomes to include and exclude in the review question
3. How to plan for identification of negative outcomes as well as the possibility of a lack of information regarding important outcomes

4. How to formulate outcomes in review questions so that they match complex social interventions which may have multi-layer objectives spanning long time frames
5. Dealing with heterogeneity of outcome questions and types
6. Correcting for inflation of outcomes evidence
7. Matching diverse reports of outcomes with the review question
8. Communication of outcome findings for responsible use by policy-makers, practitioners and service users

Which outcomes constructs to include and exclude in the review question

An immediate question that potential reviewers face is what outcomes to include and exclude in a planned review. The easy resolution is to do a quick review of the literature to determine what outcomes have been included in a sample of studies that have examined the intervention of interest. This would quickly unearth the needed information. However, this would not be a good idea. There are many reasons for this but I will mention only two so as to derail those who find this track appealing. First, it does not follow that what researchers in primary studies have selected as outcomes are important or relevant to the review question at hand. Indeed, it is likely that in some cases outcomes have been selected in primary research studies because they are favored by those conducting the research rather than those policy-makers, practitioners or service users needing answers to important questions. Second, it is also likely that in many cases researchers have selected outcomes that are easily measured within the constraints of study resources. So, while it is useful to determine what outcomes have typically been

examined in previous research prior to writing a protocol, this should not be the only or even the primary basis for selecting outcomes to include in the review question.

The Cochrane *Handbook* states that all important outcomes should be included in reviews, and that trivial outcomes should be excluded. The *Handbook* (Alderson, et al., 2004) states:

> Reviewers need to avoid overwhelming readers with data that is of little or no importance. At the same time that they must be careful not to leave out important data. If explicit criteria are necessary for establishing the presence of those outcomes these should be specified. --- Reviews should address outcomes that are meaningful to people making decisions about healthcare; it is not helpful to focus on trivial outcomes simply because those are what researchers have chosen to measure (Section 4).

A related issue pertains to the trap of focusing on small and transient outcomes. The question is not just whether an intervention has been shown to initially result in statistically significant outcomes but, more importantly, whether the intervention resulted in important and lasting improvement. That is, can important gains be generalized and maintained over time? As Gellis and Reid (2004) observe 'Frequently the main benefits seem to occur as part of the initial response to treatment and even these may be modest' (p. 172). Following their review of a number of meta-analyses covering a large variety of problems Karoly and Wheeler-Anderson (2000) conclude: '[T]reatment gains for complex and chronic problems cannot be expected to persist' (p. 172).

The first challenge to those planning systematic reviews is deciding what outcomes to include and which outcomes to exclude in the review question. How can

reviewers identify and incorporate conceptually and pragmatically meaningful outcomes as they plan their reviews? How can reviewers avoid the rush to easily measured outcomes at the expense of more complex, substantive outcomes of practical importance to policy makers, practitioners and service users? Are there some outcomes that defy measurement and, if so, does their exclusion skew or bias systematic reviews? Since there can be discrepancies between intended outcomes and outcomes that are actually measured when primary studies are implemented, how should reviewers address this potential discrepancy?

The *Handbook* states that reviewers should include all reported outcomes that are likely to be meaningful to people making a decision about the healthcare problem the review addresses. However, this begs the question of how many outcomes to include in a single review. Reviewers are urged to select outcomes that are important to decision-makers such as policy-makers and practitioners. How can a reviewer determine this in a systematic way? Reviewers are challenged to justify the specific outcomes selected as well as those to be excluded, especially when there are many possible ones to choose. A related question is whether to include outcomes that are implicit as well as those that are explicit (e.g., interventions with the explicit intended outcome of rehabilitation but with the implicit outcome that they keep juvenile delinquents “off the streets”).

Which measures of outcomes to include and exclude in the review question

Reviewers need to address a range of questions pertaining to which measures to include for the outcomes constructs selected. Reviewers need to specify not only the outcome construct but also the types of measures that will be acceptable.

Will only standardized instruments with acceptable psychometric properties be included? If so, will the reviewer specify in the protocol which standardized measures will be acceptable? If the reviewer is not familiar with many of the studies in the proposed area he or she may not be aware of which standardized measures to anticipate. Will a search be conducted of available standardized measures for the constructs of interest such as by visiting the Buros Institute's Center for Testing (at: http://www.unl.edu/buros/). If standardized measures are to be used the reviewer will need to gather information about how they were normed and the relevance of these norms to the population of interest in the proposed review. Oftentimes measured are normed on populations which are quite different from those to be examined in the proposed systematic review. Often primary researchers make modifications of standardized instruments so as to fit a particular study sample, but fail to report details of such modifications. How will modifications be systematically assessed?

Will outcome data from recordings by program personnel or employers and self-reports be acceptable and appropriate? Will self-reported outcomes be acceptable as is, or only if there are collateral reports as well? Will self-reports be treated differently?

Issues pertaining to whether or not outcomes measures will be included if those providing the data are or are not blinded as to the intervention condition becomes a concern when assessing the quality of the outcome data. Nevertheless, at the problem formulation stage the reviewers position on this measurement issue should be discussed. Most outcome studies in our fields do not obtain blinded reports, but it is probably important to distinguish unblinded data that are collected by program staff versus external researchers in terms of trustworthiness.

Is it most appropriate to use post intervention measures or change scores (difference between post- and pre scores). In addition to the use of outcomes measures will the reviewers include process measures? If so, for what purpose are process measures to be included? Will some process measures be considered indicators of immediate or intermediate outcomes?

When primary studies include multiple measures of the same construct how will reviewers determine which to use or what combination to use? How can the use of inappropriate measures of a construct be systematically detected? For example, self-esteem and self-concept are different constructs with different meanings, but an investigator may have used a measure of self-concept to measure self-esteem. How will such mistakes be detected and dealt with?

In the C2 editorial process the reviewers plans for inclusion and exclusion of outcomes measures is a matter of concern. Even at the title submission stage reviewers are expected to identify the outcome measures that will be included. The more specific reviewers can be the better.

How to plan for identification of negative outcomes as well as the possibility of a lack of information regarding important outcomes

The Cochrane *Handbook* states that reviewers should describe how they will deal with a lack of information about outcomes as well as how they will measure adverse outcomes. The *Handbook* (Alderson, et al., 2004) notes that:

> It is important to let people know when there is no reliable evidence, or no evidence about particular outcomes that are likely to be important to decision makers. --- (I)t may be important to specify outcomes that are important to decision makers, even when it is unlikely that data will be found. For example, quality of life is an

important outcome, perhaps the most important outcome, for people considering whether or not to use chemotherapy for advanced cancer, even if the available studies only report survival data. --- In addition, reviewers should indicate how they will try to include data on adverse effects in their review. In regard to this, rather than including an exhaustive list of adverse outcomes it may be more informative to summarise 'severe' (e.g. severe enough to require withdrawal of treatment) and minor adverse outcomes and include appropriate description of these (Section 3.2).

Accordingly, a challenge to reviewers is to develop a systematic plan for addressing adverse outcomes as well as the likely absence of evidence regarding an outcome.

How to formulate outcomes in review questions so that they match complex social interventions which may have multi-layer objectives spanning long time frames

It is typically thought that interventions are designed with clear objectives in mind. Sometimes these objectives are arranged in program models with sequential chains of immediate, intermediate and ultimate objectives (Chapel, 2004). These program models may have specific outcomes for each level of objective with outcome criteria and indicators attached. In complex social programs these outcomes can span considerable periods of time with more immediate outcomes occurring in the short term but ultimate outcomes occurring at distant points in time. How can reviewers formulate review questions and specify outcomes so as to systematically assess such complex intervention models? What level of objectives is to be selected and what level of outcomes should be selected? Many social program interventions are quite complex in this regard. How can a

reviewer address the complexity of such interventions in a single systematic review – or, should a series of reviews be planned in such cases?

Dealing with heterogeneity of outcome questions and types

In evaluation research it is generally thought that intervention outcomes can be of different types. For example it is common to talk about outcomes pertaining to questions of intervention *efficacy*, or *effectiveness*, or *efficiency*, or *quality*, or *equity* (Mullen, 2004). Also, in evaluation research it is common to distinguish among types of outcomes measures which may be more or less difficult to measure. For example in health care, outcomes could be assessed by using mortality indicators; physiologic indicators; clinical events; generic or specific health related quality of life measures of symptoms, of function, of care experience; or, composite measures of outcomes and time, such as disability adjusted life years (Mullen, 2004). Clearly, these outcomes address different questions and they also vary in ease of measurement.

How can reviewers consider these options in a systematic and relevant way? Especially problematic is the relative absence of findings pertaining to outcomes coming from *effectiveness* studies. Reviewers need to be especially sensitive and explicit about whether they are interested in outcomes resulting from *efficacy* or *effectiveness* studies. Randomized controlled trials are often conducted in highly controlled, artificial contexts so as to enhance internal validity (*efficacy* studies). However, such studies do not address directly how effective such interventions would be in real world contexts, such as in social agencies.

As noted by Gellis and Reid (2004):

> Client samples used in validation studies may differ substantially from the clientele in a typical agency program. The latter may be less well motivated, have a different demographic profile, and more likely to have multiple diagnoses. Will an intervention for depression tested with a sample from which clients with substance problems were excluded be effective with an agency client who suffers from both depression and alcoholism? --- experimental tests of an intervention are conducted in a supportive or at least tolerant organizational environment. Their applications in agency programs may need to be carried out in a less welcoming environment, especially if they are forced 'top-down' on staff who don't particularly want them. Transportability barriers may be especially troublesome in social work where clinical programs are multidimensional, incorporate a broad definition of service, and a multi-modal ecological approach to intervention. Such programs may generate a diversity of practice goals that may not be consistent with a goal of standardized intervention. Such considerations suggest that interventions tested in a RCT may not necessarily be applied in the same way or with the same degree of efficacy when they reach the level of agency practice" (p. 160).

When interventions are conceptualized at a theoretical level it may not be necessary to take into consideration intervention context. However, when interventions are implemented and evaluated, it is generally recognized that associated outcomes will be context dependent. How can reviewers address this context dependency when selecting and defining outcomes for inclusion in a systematic review? Should these

distinctions matter to reviewers when selecting outcomes to include in a review? If so, how can these challenges be addressed?

Inflation of evidence

Inflation of evidence can result from many sources including the well-known tendency for investigators or their assistants to give ‘a leg up’ to interventions they may have helped to develop or that for some reason they favor. It is also possible that investigators will propose and measure outcomes that are known to favour their interventions while avoiding those not favouring their intervention. Inflation of evidence can also result from selection of client self-report to measure outcomes and that such reports may reflect social desirability, expectancy, or cognitive dissonance effects (Gellis and Reid, p. 159).

This issue is related to detection bias. The *Handbook* (Alderson, et al., 2004) refers to detection bias as follows:

> Detection bias refers to systematic differences between the comparison groups in outcome assessment. Trials that blind the people who will assess outcomes to the intervention allocation should logically be less likely to be biased than trials that do not. Blinding is likely to be particularly important in research with subjective outcome measures such as pain --- Bias due to the selective reporting of results is somewhat different from bias in outcome assessment. This source of bias may be important in areas where multiple outcome measures are used ---. Therefore, reviewers may want to consider specification of predefined primary outcomes and analyses by the investigators as indicators of validity. Alternatively, selective

reporting of particular outcomes could be taken to suggest the need for better reporting and efforts by reviewers to obtain missing data. (Section 6.6)

The challenge here is to determine how reviewers can avoid inflation of evidence bias when selecting outcomes for inclusion in reviews.

Matching diverse reports of outcomes with the review question

Having succeeded in specifying in the review question which outcomes will be of interest, reviewers may find that studies include many other outcomes of interest. As noted in the *Handbook* (Alderson, et al., 2004*)*:

> Reports of studies often include more than one outcome (mortality, morbidity, quality of life, etc.), may report the same outcome using different measures, may include outcomes for subgroups and may report outcomes measured at different points in time. The reviewer needs to integrate what type of outcome information is needed to answer the review's question(s) with what is likely to be in the reports of studies (Section 7.5.4).

The challenge here is to prepare for this problem and to include in the protocol a plan for matching diverse reports of outcomes with the review question so as not to be distracted by the possible avalanche of irrelevant information that may be found in primary studies.

Reporting findings

The *Handbook* discusses some of the issues associated with drawing conclusions about outcomes found in a systematic review. The *Handbook* (Alderson, et al., 2004) states:

> In addition to considering the strength of evidence underlying any conclusions that are drawn, reviewers should be as explicit as possible about any judgements about preferences (the values attached to different outcomes) that they make. Healthcare interventions generally entail costs and risks of harm, as well as expectations of benefit. Drawing conclusions about the practical usefulness of an intervention entails making trade-offs, either implicitly or explicitly, between the estimated benefits and the estimated costs and harms --- reviewers should consider all of the potentially important outcomes of an intervention when drawing conclusions, including ones for which there may be no reliable data from the included trials. They should also be cautious about any assumptions they make about the relative value of the benefits, harms and costs of an intervention (Section 9.5).

The challenge then is to effectively communicate what is found regarding outcomes but also to find ways of placing those findings in the context of other important considerations that decision makers need to consider. The relative value of the outcomes, including benefits and harms, as well as costs are among the more important considerations.

A somewhat separate issue regarding the communication and use of review findings pertaining to outcomes has to do with the importance of recognizing and communicating variation in outcomes within intervention or treatment groups. As noted by Mullen and Streiner (2004):

> The results of RCTs are analyzed by comparing the mean score of the experimental group against that of the placebo or control group (or some comparable

> summary statistic). However, this masks the fact that there is always individual variability around the means, and overlap in the distributions of scores for the two groups. The result of this is that a proportion of people in the experimental group actually do worse than some in the control group and, conversely, some in the comparison group improve more than some people in the active treatment group. The implication is that practitioners cannot blindly apply a 'proven' procedure and assume that a particular individual receiving that procedure will benefit ---. This has led some critics to reject the whole notion of (evidence-based policy and practice), stating that results of trials are incapable of being applied at the level of the individual (pp. 115-116).

When communicating the results of systematic reviews this issue gets even more complex. The challenge here is to communicate findings about outcomes across reviewed studies so as to make clear that average differences do not apply to individual cases and that variation is to be expected.

These are a few challenges which I think reviewers must face as they consider outcomes measurement. In keeping with the idea of a systematic review, the protocol should make explicit how the reviewer plans to deal with each of these challenges.

I have limited my comments to systematic reviews in general and I have not dipped into special challenges that face those contemplating meta-analyses. I expect there may be some interesting additional issues facing those planning to use meta-analysis. I look forward to Professor Melhuish's discussion and to our subsequent discussion about these and other challenges of relevance to outcomes measurement in systematic reviews.

References

Alderson P, Green S, Higgins JPT, editors. *Cochrane Reviewers' Handbook 4.2.2* [updated December 2003]. In: The Cochrane Library, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.

Australia Institute of Health and Welfare (2000). *Australia's Health 2000*. Canberra: Author.

Chapel, T. (2004). Constructing and using logic models in program evaluation. In A. R. Roberts & K. R. Yeager (Eds.), *Evidence-based Practice Manual: Research and Outcome Measures in Health and Human Services*. Oxford: Oxford University Press.

Donabedian, A. (1981). Criteria, norms and standards of quality–what do they mean. *American Journal of Public Health, 71*(4), 409-412.

Gellis, Z., & Reid, W. J. (2004). Strengthening evidence-based practice. *Brief Treatment and Crisis Intervention, 4*(2).

Government Performance and Results Act of 1993, Pub. L. No. 103-62; § 1115, 107 Stat. 285. (1993).

Karoly, P. & Wheeler-Anderson, C. (2000). The long and short of psychological change: Toward a goal-centered understanding of treatment durability and adaptive success. In C.R. Snyder and R.E. Ingram (Eds.). *Handbook of Psychological Change. (pp. 154-176)* New York: John Wiley and Sons, Inc.

Mullen, E. J. (2004). Outcomes measurement: A social work framework for health and mental health. *Social Work in Mental Health, 2*(2).

Also published in: Mullen, E. J. (2004). *Evidence-based Practice in a Social Work Context - The United States Case* (Vol. FinSoc Working Papers 2/04). Helsinki, Finland: STAKES. http://www.stakes.fi/finsoc/mullen-evidence-based_practice.pdf

Mullen, E. J., & Magnabosco, J. L. (1997). *Outcomes Measurement in the Human services: Cross-cutting Issues and Methods*. Washington, DC: NASW Press. http://www.naswpress.org/publications/books/clinical/outcomes_measurement/2758.html

Mullen, E. J., & Streiner, D. L. (2004). The evidence for and against evidence based practice. *Brief Treatment and Crisis Intervention, 4*(2).

Rossi, P. H. (1997). Program outcomes: Conceptual and measurement issues. In E. J. Mullen & J. l. Magnabosco (Eds.), *Outcomes Measurement in the Human Services*. Washington, D.C.: NASW Press.

United Kingdom Clearing House on Health Outcomes. (March 1997). *Definitions of Outcomes*. Retrieved July 26, 2002, http://www.leeds.ac.uk/nuffield/infoservices/UKCH/define.html.