Responsible Development in a Polycentric World
Inequality, Citizenship and the Middle Classes
14th EADI General Conference
23-26 June 2014, Bonn
www.gc2014.org

How to Choose Methods for Evaluation Research: A Test of the Relevance of Systematic Review

Author: Aogán Delaney;
Peter Tamas;
Hilde Tobi

Institution: Research Methodology, Wageningen University

Address: Postbus 8130
6700 EW Wageningen

Email: aogan.delaney@gmail.com

Abstract

Systematic review is used when authors want a comprehensive understanding of what is known on a given topic. It is the gold standard in the health sciences and it is of increasing importance in the social sciences. Systematic reviews are typically used to identify, assess and synthesize empirical evidence. This paper reports on a test of the suitability of systematic review for the identification of research methods appropriate for evaluation research on the effects of primary health care delivery on peace and conflict dynamics in conflict-affected regions. Our review methodology involved a reproducible search strategy, inclusion protocols for identification of relevant articles, criteria for quality assessment, defined procedures for extraction of data, and a systematic aggregation method. Testing this design finds that adopting even some aspects of systematic review are an improvement over the academic methodologically prescriptive articles reviewed. However, proper use of systematic review is complicated by the paucity of empirically grounded methodological recommendations, the prevalence of non-empirical arguments, uncertain citation practices and inconsistent reporting in the articles reviewed.

Keywords

developing countries; research design; systematic review; community and public health; conflict studies

As suggested by the theme of this working group, there is a growing consensus around the need for evaluations of development projects and policies that are valid, reliable and relevant. Equally, the theme suggests that it is difficult to determine which interventions work and which do not. This paper reports on a partially successful attempt to use systematic review (SR) to identify appropriate evaluation methods.

The search for 'what works' and evidence-based policy can be seen as a response to some of the successes resulting from the evidence-based-medicine movement in the medical and health sciences. The successes in the health sciences came from thorough testing of interventions through Randomised Controlled Trials (RCTs) and subsequent meta-analysis of RCTs through systematic review. (Dixon-Woods, Bonas, et al. 2006; Evans and Benefield 2001; Magarey 2001; Major and Savin-Baden 2012). While the successes are impressive, there are differences which limit the extent to which this philosophy can be transferred to the study of development. Among the most obvious methodological constraints is the inability to account for external contextual factors (Roberts et al. 2010), ethical dilemmas that come with suggesting the use of 'placebo' development interventions as control groups (Gutlove and Thompson 2006), and, at a theoretical level, limited inter-disciplinary consensus with respect to constructs, indicators and phenomena (C. Buhmann et al. 2010).

While not mentioned as frequently, an additional limitation is that evaluators come from academic disciplines and every academic discipline has its own, at times deeply contested, diversity of theories, methods and standards. This diversity eliminates the possibility for interdisciplinary criteria by which to judge the quality of these methods (Weaver and Roberts 2010; C. Buhmann et al. 2010). The lack of inter-disciplinary standards for the assessment of research methods points to a knowledge gap that limits development of sound inter-disciplinary evaluation methods suitable for the examination of development programmes. In order for an impact evaluation to produce results that are considered valid by all concerned disciplinary fields, the evaluators will first need to construct a design that satisfies the methodological quality standards found in all of those disciplines. However, because methodological discussion and innovation tends to happen within communities where mutual understanding is possible, that is within disciplines, and because researchers tend to be formed within one of these communities, researchers are not well equipped when they are required to negotiate cross-disciplinary benchmarks with counterparts from other disciplines, counterparts who may have very different notions regarding sound research design. Or, to put it in layman's' terms, in order to find out what works, we first need to find out what works about the methods we use to find out what works. This paper reports on what we did when asked by a non-governmental organization (NGO) to work across disciplines to identify research methods that work.

In early 2012 an international NGO asked our research methodology group at Wageningen University for advice on how to study the interaction between the manner in which they organized the delivery of primary health care services and peace and conflict dynamics in a province of Afghanistan. They were the sole provider of primary healthcare in this area of endemic conflict and continued funding of their activities depended on demonstrating a relationship between service-delivery and levels of peace and conflict in the region. In discussion with the NGO and an expert in the field we were informed that debates over the relevance of the organization of primary health care provision in areas of endemic conflict are highly politicized. The social relevance and politicization of the topic discouraged us from giving expert advice as that form of review has well known deficiencies (e.g. Antman 1992; Schaafsma et al. 2005). We, instead, chose to use this opportunity to test the relevance of systematic review for the task of generating methodological prescriptions from a heterogeneous inter-disciplinary literature.

The proposed impact evaluation study would fall between the health sciences and the multi-disciplinary field of conflict studies. The interaction between health care and conflict levels is an emerging field of study, but one that lacks an adequate evidence base (Bornemisza et al. 2010; C. B. Buhmann 2005; Gutlove and Thompson 2006; MacQueen and Santa-Barbara 2000), due in part to the small number of empirical studies undertaken and to the wide range of conceptual frameworks and methodological approaches used, making meaningful comparison of studies difficult (C. B. Buhmann 2005; Roberts et al. 2010).

Although a tempting solution would have been to draw upon the best practices in one of the relevant disciplines, and establish that as the standard for research in this emerging field, we foresaw two principal problems with this strategy. First, prioritising the methods of one discipline would accentuate the biases of that discipline and limit the insights from the other disciplinary approaches Second, we thought that the very selection of one discipline by expert consultants would constitute a regressive shift away from empiricism in favour of expert opinion, and would constitute a further politicization of an already sensitive topic.

If a solution was to be found to this dilemma, we thought it might lie in using the best practices from *all* relevant disciplines, and in assembling compatible best practices through scientific methods as opposed to expert selection. We therefore hypothesised that using a systematic review to generate best practices for designing research into the relation between health care delivery and dynamics of peace and conflict would yield a stronger research design than either the state of the art in one discipline, or the expert opinion of academics.

This paper is structured as follows. We briefly discuss the evolution of the Systematic Review methodology and the expanding number of application areas. From this discussion we formulate three research questions that guided our test of the SR methodology. In the third section we describe how our methodology was designed and implemented. Afterwards we report

on how successful we were with this research design before discussing these relative benefits and disadvantages and drawing conclusions and recommendations for future directions of research.

Theoretical framework & research questions

Systematic review has been adopted and adapted extensively in the health sciences to survey broad empirical literature. In the social sciences, SRs are now also used to make sense of and manage the 'information explosion' (Major and Savin-Baden 2012; Wallace et al. 2004), separate wheat from the chaff (Major and Savin-Baden 2012; Wallace et al. 2004), identify gaps in an evidence base (Major and Savin-Baden 2012; Wallace et al. 2004), confirm, refute, develop, or modify bodies of theory (Campbell et al. 2003; Noblit and Hare 1988), and increase the standard of research in the field (Bondas and Hall 2007). SRs are used in education and training (Evans and Benefield 2001; Price 2005; Secomb 2008), social policy (Wallace et al. 2004), and qualitative research of patients' experiences of medical conditions (for example Arman and Rehnsfeldt 2003; Campbell et al. 2003; Dixon-Woods, Cavers, et al. 2006; Hughes, Closs, and Clark 2009; Dixon-Woods et al. 2007). The expanding range of domains in which systematic review is applied and its suitability for working from a broad literature commended it to our use. Further, to our knowledge, the relevance of systematic review to the generation of methodological prescriptions for interdisciplinary research has not been explicitly tested.

Systematic Reviews were used originally to conduct a meta-analysis of all known RCT studies investigating a relationship between a given intervention and a desired outcome (Dixon-Woods, Bonas, et al. 2006, -; Magarey 2001). The methodology is composed of the following steps: transparent and reproducible search strategy; selection of studies to be included in the review; data extraction; secondary analysis of extracted data. Each of these steps is pursued through a method that embodies the principles of transparency, reliability, and comprehensiveness. As SR has been adapted to cover topics in which the goals of research, types of evidence, field conditions, and epistemological foundations of the health and medical sciences no longer hold, each of these methods has required adaptation, although in such a way as to remain committed to the underlying principles of transparency, reliability, and comprehensiveness. In order to test our working hypothesis, we structured a set of three research questions around the transfer of these components to a review of methodological prescriptions. These Research Questions are formulated as follows:

Given the task of generating prescriptions for an inter-disciplinary study of the interaction between the organization of primary health care provision and peace and conflict dynamics in an area of endemic conflict from a heterogeneous methodologically prescriptive literature:

1. What aspects of systematic review transfer well?
2. What aspects of systematic review do not transfer well?
3. Are the aspects of systematic review tested an improvement on current practice?

Methods

The review we undertook followed the typical structure of a systematic review as we have found them in the social sciences. Specifically, we:

1. conducted a recorded search for articles related to the topic
2. narrowed our population of articles down to those that were written in English
3. screened the titles and abstracts of articles according to standardised protocols for relevance
4. screened the full text of retrieved articles for relevance, according to standardised protocols
5. assessed the quality of the articles found relevant
6. identified and extracted analytically relevant data within these articles
7. summarized analytically relevant data

Systematic review requires clear specification of each of these steps prior to starting a study. We did not pre-specify all steps as this was a test application of systematic review to a new task. As such, for each stage we identified challenges, picked a path and then documented our progress. Each of these stages is discussed in summary form below.

Search

Systematic review requires identification and review of material in obscure sources and, as such, it is enormously labour intensive. In our project, as seems common in systematic reviews in the social sciences, we accepted publication in a refereed journal as an initial screen for quality and limited our search to iterative creation of a complex search term that we executed in indexes of these refereed journals.

This search term was recorded and is included at the end of this article (Figure 1 and Figure 2) along with all protocols mentioned in the remainder of this section. After duplicates were removed, this search terms yielded 312 articles, which we used as our population in the analysis that followed.

Screen for language

The primary researcher could only read in English. As such, and while there are likely serious problems with this, we rejected non-English articles, reducing our set of articles to 269.

Screen for relevance

Once we had downloaded the titles and abstracts and eliminated duplicates, we created and applied a standardized protocol (Figure 3) to perform an initial screen for relevance based on article titles and abstracts. This stage of screening resulted with 168 possibly relevant, English-language articles.

We created a standardized protocol (Figure 4) to screen the full text of articles for relevance. We used this to screen 162 of the 168 preliminary relevant articles. 6 articles could not be screened as we could not access the full text of the article, neither through our library, nor through direct correspondence with the authors. This stage of screening resulted in 64 relevant articles.

Assess quality

Quality assessment is usually done through the application of accepted standards to a hopefully adequate discussion of the methods used in independent empirical studies. Our retrieved articles were a mix of empirical studies and non-empirical arguments that were informed by a diversity of theoretical perspectives. As such, a single external quality standard could never be fair. We, therefore, operationalized and tried to assess articles according to a cross-disciplinary standard of 'internal coherence' (Figure 5).

Identify analytically relevant data

We developed a coding scheme that we first tested and then applied, top down, to all retained articles that identified and categorised analytically relevant data.

Analyse data

The analysis stage in systematic review involves summarizing source material. Within methodological literature on synthesis of qualitative research, there is a useful heuristic distinction between interpretive and aggregative synthesis strategies (Noblit and Hare 1988; Dixon-Woods, Cavers, et al. 2006). Translated to our analysis, we were confronted with a trade-off between elegance and adequacy. Reducing heterogeneous source material to a single elegant

narrative, an interpretive synthesis, involves both the destruction of data and forms of creativity that are hard to make transparent and reproducible. For this review we tried to aggregate, that is, to retain all underlying data and to process these data in a transparent and reproducible manner.

We tried to aggregate data twice. In our first try the methodological prescriptions we identified included many duplications with relevance across a large number of thematic categories. To structure these prescriptions we tried to identify the stage of research (i.e. design to reporting), the study design (e.g. cross-sectional) and the nature of the research context (e.g. a war zone) to which they were relevant. These categories did not support adequately structuring the data and, while the resulting synthesis was useful to the NGO, it did not meet our own standards.

Our second attempt at aggregation began with a return to the literature in order to identify an appropriate alternative. There has been significant work on qualitative synthesis methodologies (Bondas and Hall 2007; Campbell et al. 2003; Dixon-Woods et al. 2007; Dixon-Woods, Cavers, et al. 2006; Higginbottom et al. 2012; Major and Savin-Baden 2012) but we considered that these approaches lack systematicity and veer either towards radical constructivism (i.e. critical interpretive synthesis, grounded theory, meta-ethnography, narrative summary, qualitative research synthesis, and thematic analysis) or to discard context in quantifying qualitative text (i.e. Bayesian meta-analysis, case survey, content analysis, and qualitative comparative analysis). With this in mind, we developed and used an aggregation method that drew on cross-case techniques (Miles and Huberman 1994), narrative summary, (Hubbard, Kidd, and Donaghy 2008; Secomb 2008), and meta-study (Paterson et al. 2001). We tried this aggregation procedure on a sub-set of articles that were most immediately relevant to the NGO.

The process by which we identified prescriptions in articles was:

1. Operationalize the research question[1] in terms of sub-questions
2. Code all articles top-down using those sub-questions
3. Code all articles bottom-up for missed analytically relevant themes and for variables that may be useful for categorizing prescriptions

---

1

This is the research question which the systematic review sought to address: "What methodological guidelines for research into the interaction between the delivery of primary health care and peace and conflict dynamics in areas of endemic conflict can be extrapolated from methodological prescriptive social science literature?" This is not to be confused with the research question of this paper which is concerned with investigating hte usefulness of the SR methodology.

4. Assess bottom up codes and recode all articles top-down based on that assessment
5. For each article, write a narrative summary accounting for all coded text, with accompanying log of analytically relevant decisions made in preparing each narrative summary
6. Code narrative summaries for analytically relevant themes
7. Aggregate coded text into thematic clusters by codes.
8. Within each cluster collapse identical statements
9. Within each cluster explain incompatible statements by referencing original articles
10. Convert remaining statements into appropriately qualified methodological prescriptions referencing original articles, log notes and explanations of divergence

Contrary to our first aggregation attempt, in our second attempt we identified all candidate categorizing variables through bottom-up coding. All candidate categorizing variables were then applied through top-down recoding of all articles. These candidates were assessed with respect to their analytic relevance, their prevalence in and the extent to which they cleanly segmented the reviewed literature.

Results

In this section we report on how successful we were with each element of our methodology.

Search and screening

We were successful in creating a recordable search strategy to generate an initial pool of articles. The scope of our review was limited by a language and publication bias. We were able to use protocols for selection of relevant articles that are standard within systematic reviews. We were unable to assess articles for quality as the heterogeneity of the returned articles destroyed our ability to identify a fair screening method and we were unable to create a reliable measure of 'internal coherence'. As has been found by other researchers (e.g. Fischer, Tobi, and Ronteltap 2011), our search returned a large fraction of non-empirical articles. We were able to identify methodological prescriptions in all source articles.

Synthesis

In working through our first attempt at aggregation we found that our pre-specified categorizing variables did not allow us to reliably identify either dependencies or (in)compatibilities between prescriptions which destroyed our ability to produce an adequate synthesis. In our second

attempt at aggregation the best categorizing variable we found was 'reach' where that was operationalized as 'how close to the Real World does this prescription get the researcher?", with possible values *real world; interpretations of the real world; contextually shaped constructions of the real world;* and *contextually shaped representations of contextually shaped constructions of the real world*. We found that the heterogeneity of source articles made it impossible to create narrative summaries that were analytically preferable to the underlying articles. We had difficulties identifying both mutual dependencies and the limit of the relevance of prescriptions but we found that 'reach' allowed us to judge the compatibility of prescriptions. We were unable to judge the quality of or weight prescriptions despite established qualitative methods such as refutational synthesis or lines-of-argument synthesis (Noblit and Hare 1988). In neither attempt at synthesis could we find a reliable means to decide when a given prescription stopped being relevant.

Discussion

Aspects that transferred well

We were able to execute a search and screen returned articles for relevance to an interdisciplinary study on the organization of health care in a comprehensive and transparent manner. While this was to be expected, as abstract and text appraisals are standard practice for determining relevance in many systematic reviews (Campbell et al. 2003; Dixon-Woods et al. 2007; Higginbottom et al. 2012; Hubbard, Kidd, and Donaghy 2008), such a transparent search for and selection of articles has been noted as wanting in reviews of argument-based literature (McCullough, Coverdale, and Chervenak 2007, 72). We were able transparently to identify, classify analyse and process prescriptions from a literature that spanned medical and social sciences in a rigorous, comprehensive and transparent manner. Each of these marks an improvement over the practice in the articles we reviewed.

Aspects that did not transfer well

Our review contained an English-language bias and a publication bias. 32 articles (10.26%) from the initial 312 returned by our search were excluded due to being written in a language other than English. As language-screening was the first refinement step, no further investigation was made as to the relevance of these 32 articles. The language bias might therefore be considered quantitatively small, but may re-enforce divergences in standards between linguistically-segregated academic communities. Our language screen was due to resource constraints, and

while not uncommon in systematic reviews (Magarey 2001), we consider this an important and relatively simple area for improvement.

We confined our search to databases of peer-reviewed academic articles. Grey-literature was therefore not included in our review. While this may exclude some important methodological innovations made for example in the practitioner literature, we don't yet see a way to overcome this publication bias as it serves two important functions: quantity and quality management. While statistical meta-analysis systematic reviews are strengthened with larger numbers of studies, with qualitative systematic reviews, due to the resources required for analysing large bodies of qualitative data, it is not uncommon for a review to cover a small but focussed sample (as an example, Campbell et al (2003) reviewed 10 studies in their meta-ethnography of the experiences of diabetics). An earlier version of our search on Web of Science returned 13,036 articles, an unworkably large number. To refine, we focussed the search at a higher level of abstraction. It is likely that extending the search to grey literature would return a similarly unworkable number of articles.

We were not able to appraise the quality of reviewed articles. We used 'indexed in Scopus or Web of Science', which externalises the problem of quality appraisal from us as reviewers to the known imperfections of inclusion in indexes of peer-reviewed sources. Further, we were unable to adequately assess internal coherence. However, we find that this is no worse than the practice found in the articles reviewed, which rarely discuss the quality of articles they cite when building methodological recommendations based on those cited articles.

In no case did the prescriptions made in a single article discuss everything methodologically required for a given research effort. The prescriptions in a given article always had external dependencies. These dependencies would, ideally, be comprehensively identifiable through citations. We, however, had no evidence that the authors whose articles we reviewed were systematic about their citations. Authors' lack of transparency in citation made it impossible for us to falsify the hypothesis that they just cherry-picked citations that made their prescriptions look good. We were, therefore, unable to properly delimit individually adequate and mutually discrete methodological prescriptions.

The articles we studied relied variously on evidence and argument to justify their prescriptions. In aggregating prescriptions arising from these very different foundations we assumed equivalence. In trying to weight prescriptions, we considered various strategies based on principles of source quality, rational completeness, and cross-disciplinary or cross-epistemological triangulation and consensus. However, it appears to us that all of these strategies would require considerable methodological development to the extent that they remain well beyond the scope of a single study. Of these, it seems that the principle of source quality seems the most promising, but is so far subject to limitations discussed elsewhere in this paper.

Another essential area for development of synthesis methods is in the integration of evidence-based and argument-based knowledge, and within evidence-based knowledge, that based on quantitative, qualitative, and mixed methods research. While some work in theorising the value of different types of knowledge to a review framework appears promising (Heyvaert, Maes, and Onghena 2011; McCullough, Coverdale, and Chervenak 2007), our experience suggests that the challenges are formidable. Specifically, such developments will need to deal with the traditional qualitative-quantitative divide, the evidence-theory divide, their relationships to forms of knowledge on the epistemological spectrum, and whether such dichotomous categories are at all justified. Although our search strategy specifically targeted articles with a conscious ontological awareness, the heterogeneity of returned articles does not give us confidence that it will be easy to find an overarching framework within which each article can be neatly placed. In the absence of a forthcoming framework, we would argue in line with the principals of systematic review and evaluation that deference be given to prescriptions supported by *evidence*. Our un-weighted aggregation of epistemologically distinct rationales was therefore unfortunate but was and will remain a challenge until such time as methodological prescriptions are only publishable when supported by evidence.

Conclusion

In this essay we found that many of the procedures standard within systematic review are relevant to the generation of methodological prescriptions from a heterogeneous literature. In our case, the additional effort required for a transparent and reproducible review was merited. Based on this experience, where at all possible, we recommend the use of a systematic approach for the design of evaluation studies in complex environments. We consider that using systematic review methods to design research will increase the quality of interdisciplinary development evaluation. In particular our success with using a recordable literature search strategy, transparent inclusion protocols, and a systematic and comprehensive aggregative framework, sets a standard for methodologists far above that observed in most of articles we reviewed. We recommend the use of these three components as a minimum when designing development evaluation studies. We expect such an approach to yield research designs that are transparent about their sensitivity to the methodological requirements arising from each of the contributing disciplines, to methodological innovations and debates within each of these disciplines, and to the inconvenient attributes of the methodologies considered.

We did encounter four frustrations that we expect will trouble those who follow our example: the articles identified by our search were sufficiently heterogeneous to disrupt comparison, many of the articles we reviewed were non-empirical and relied on argument rather

than evidence, we could not find cross-disciplinary standards by which we could assess the quality of articles, and authors were not transparent in their own citation of supporting work.

Two of the frustrations we encountered are being worked on. Scholars are extending systematic review to heterogeneous literature. In this regard critical interpretive synthesis (Dixon-Woods, Cavers, et al. 2006) and Heyvaert, Maes, and Onghena's (2011) proposed classification of 18 mixed methods research synthesis frameworks are exemplary. This effort invites further development, although its proximity to fundamental debates and assumptions in epistemological categorisation cautions us against expectations of any quick-fix solution. In our review we were confronted by prescriptions that were justified by reference to argument. While there is some work on improving the review of argument-based literature (e.g. Mahieu and Gastmans 2012; McCullough, Coverdale, and Chervenak 2007; Sofaer and Strech 2012; Sofaer and Strech 2011), our hope is that it becomes impossible to publish methodological prescriptions without an evidentiary basis.

This research was motivated partly in response to a difficulty in designing interdisciplinary research due to the plurality of quality standards within each relevant discipline. Instead of overcoming this issue we instead were confronted with the same problem at a higher level of abstraction: it was impossible to screen the methodologically prescriptive articles we reviewed on quality grounds because we could not identify acceptable cross-disciplinary standards by which to assess articles. While the evidence regarding peer-review suggests that it is a poor proxy for quality appraisals, and while an academic-publication bias may exclude some relevant innovations arising from 'grey' evaluation studies, this seems to be the only reasonable minimum stop-gap standard. We hope a systematic approach to reviews will create pressure for the identification of cross-disciplinary quality criteria.

The final problem we encountered was that the authors we read did not discuss how they picked their source articles. Transparency of reporting is, in some sectors, used as a proxy for quality. This appears as 'signal to noise' (Dixon-Woods, Cavers, et al. 2006; Edwards, Russell, and Stott 1998; Hughes, Closs, and Clark 2009), 'credibility' (Atkins et al. 2012), and 'adequate reporting of methods' (Carroll, Booth, and Lloyd-Jones 2012). While a poor proxy for quality at best, such transparency would help reviewers identify both the dependencies of prescriptions within their supporting literature and the bounds of their relevance and, as such, would seem reasonable for editors to require of their authors.

As a consequence of these four frustrations, heterogeneity, non-empirical argument, the absence of cross-disciplinary standards, and the absence of transparency in citations, we were ultimately unable either to weight our prescriptions or to deal with conflicting standards. Therefore we welcome efforts by methodologists to overcome these four obstacles, and we encourage further use of systematic approaches to review in order to create pressure for such

methodological developments. We believe improvements in systematic reviews along these four lines will do much to raise standards of evaluation research through providing a means through which a cross-disciplinary consensus can be reached on methodological priorities for impact evaluation research in complex development environments.

Appendices

Figure 1: Details of search executed in Web of Science

#	Hits	Search details	
# 5	124	(#4 and (ts=(epist*) or ti=(epist*))) AND Document Types=(Article) Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=All Years Lemmatization=On	Find all those articles in #4 that indicate an interest in epistemology.
# 4	13,036	#3 OR #2 Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=All Years Lemmatization=On	Combine #3 and #2
# 3	11,128	(TI=((methodo* OR research OR knowledge OR real* OR post-* OR constructivi* OR neo-positiv* OR interpretiv* OR emotionali* OR emic OR subjectivi*) AND (quality OR standard* OR rigor OR rigour OR fidelity OR criteria OR valid* OR judg* OR metro* OR reliab*))) AND Document Types=(Article) Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=All Years Lemmatization=On	identify those articles whose titles indicated both an interest in methodology or research or a theoretical perspective and an interest in quality
# 2	3,018	(#1 and TI=(methodo* OR research OR knowledge) and TS=(social or cultur* or policy or applied)) AND Document Types=(Article) Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=All Years Lemmatization=On	Find all of those articles from the first set which have words in the title indicating an interest in methodology and words in the title or abstract indicating an interest in social phenomena or application or policy relevance
# 1	80,450	(TS=((methodo* OR research OR knowledge OR real* OR post-* OR constructivi* OR neo-positiv* OR interpretiv* OR emotionali* OR emic OR	Find all versions of words indicating an interest in knowledge or methodology or

		subjectivi*) NEAR/5 (quality OR standard* OR rigor OR rigour OR fidelity OR criteria OR valid* OR judg* OR metro* OR reliab*))) AND Document Types=(Article) Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH Timespan=All Years Lemmatization=On	a theoretical perspective within five words of a word that references an interest in quality.

Figure 2: Details of Search term executed in Scopus

((TITLE-ABS-KEY(epist*)) AND (TITLE(methodo* OR research OR knowledge OR reali* OR post-mo* OR post-po* OR post-str* OR constructivi* OR neo-positiv* OR interpretiv* OR emotionali* OR emic OR subjectivi*) AND TITLE(quality OR standard* OR rigor OR rigour OR fidelity OR criteria OR valid* OR judg* OR metro* OR reliab*))) OR ((TITLE(methodo* OR research OR knowledge) AND ABS(social or cultur* or policy or applied)) AND (((TITLE-ABS-KEY(epist*)) AND (ABS(quality OR standard* OR rigor OR rigour OR fidelity OR criteria OR valid* OR judg* OR metro* OR reliab*))) AND ((ABS(methodo* OR research OR knowledge)) OR (ABS(reali*)) OR (ABS(constructivi* OR neo-positiv* OR interpretiv* OR emotionali* OR emic OR subjectivi*)) OR (ABS(post-st* OR post-mo* OR post-po*)))))

Figure 3: Relevance-screening Protocol for article title and abstract

$C_{RELEVANT}$: The abstract states that the article advocates or problematizes some particular methodological standards relevant to any epistemological orientation or any stage of research

Figure 4: Relevance-screening Protocols for full article

Criteria	**Rationale**
$C_{FULLTEXT\ 1}$: The standards advocated/problematized by the article must be applicable to the applied social science field of health	Our search strategy produced a set of articles from a variety of different social science and applied social science fields of study. Many of these would not be relevant for the funder's research purposes. This criterion

and peace/conflict.	sought to include only articles whose prescriptions would be relevant for research in health and/or conflict studies.
$C_{FULLTEXT\ 2}$: The standards advocated/problematized by the article must be applicable to one or more of the following stages of research: Refinement; Research; Analysis; Interpretation; Evaluation.	A number of the articles identified by the search strategy were concerned with prescribing standards for commissioning research, while others were concerned with standards for publication and dissemination of results of research. We considered both of these stages of research to be unhelpful for the purposes of the funder's research. We therefore designed this criterion in order to select only articles that prescribe standards for a relevant stage of research.
$C_{FULLTEXT\ 3}$: The standards advocated/problematized by the article must be applicable to research within a realist or positivist epistemology.	The initial search strategy was epistemologically open. This therefore produced a set of articles prescribing standards for research across the epistemological spectrum. The intended purpose of the funder's research was to investigate possible relationships between delivery of primary health care and patterns of peace and conflict dynamics, a problem that is located in a critical realist tradition of research. We designed this criterion to exclude articles that did not make prescriptions applicable to critical realist research.

Figure 5: Protocols for assessing article quality ('internal coherence')

Criteria	**Rationale**
$C_{COHERENT\ 1}$: Any knowledge claims, whether resulting from original research, or from expanding on or critiquing existing work, or as a result of theoretical and/or abstract argument, must be formulated within the epistemological	The papers we encountered in this review were from a variety of different forms and made knowledge claims on a variety of different bases, namely argument based theorising, responding to established work, and original empirical research. In an effort to apply a standard screening criterion that would work across these different forms, we considered that a basic feature of internal

orientation of the article.	coherence was that each article would be epistemologically consistent. We formulated this criterion to enquire whether the knowledge-claims of an article are epistemologically consistent with the article.
$C_{COHERENT\ 2}$: If the knowledge claims of the paper are based on original research, and if that research falls within a research stage or epistemic orientation to which the knowledge claims advocate standards, then that research must adhere to those standards.	For those empirical articles that met the first criterion of epistemological consistency, we devised this second criterion to evaluate the quality of research based on the standards that the authors themselves advocate. We considered that this criterion could only be applicable in cases where the type of research upon which an article is based would be applicable to the prescriptions advocated in the article.
$C_{COHERENT\ 3}$: If the knowledge claims of the paper are based on original research of an epistemic orientation or stage other than those that the knowledge claims advocate standards for, then research must adhere to previously accepted standards, unless those standards have been problematized in the paper.	For those empirical articles that were deemed to be epistemologically consistent, but which prescribed standards for a type of research other than that on which the article was based, we formulated this criterion which appeals to previously accepted standards of quality.
$C_{COHERENT\ 4}$: If the knowledge claims of the paper are produced based on expanding existing work, then the knowledge claims must reflect the essence (in terms of epistemology and substantive argument) of the existing work and the context within which the existing work is expanded on, unless the existing work or context is critiqued.	These final two criteria were formulated in order to establish quality standards for argument-based articles, based on argumentative and logical consistency. Criterion 4 deals with articles that build their knowledge-claims on expanding previous works, and establishes as a basic standard of quality that the conclusions of the article should be consistent with the logic of the previous work and with the particular argumentative contribution being made in the article, with the exception of when previous work is critiqued. Criterion 5 deals with articles that build knowledge-claims based on critiques. It establishes a basic principal of argumentative consistency.
CCOHERENT 5: If the case for the	

knowledge claims of the paper is made based on a critique of existing work or the context of its application, or based on theoretical and/or abstract argumentation, then the case for knowledge claims must be made on the same basis as the critique or theoretical/abstract argumentation.	

Bibliography

Antman, E M. 1992. 'A Comparison of Results of Meta-Analyses of Randomized Control Trials and Recommendations of Clinical Experts. Treatments for Myocardial Infarction'. Edited by J Lau, B Kupelnick, F Mosteller, and T C Chalmers. *JAMA : The Journal of the American Medical Association, 1992, Vol.268(2), pp.240-8* 268 (2): 240–48.

Arman, Maria, and Arne Rehnsfeldt. 2003. 'The Hidden Suffering Among Breast Cancer Patients: A Qualitative Metasynthesis'. *Qualitative Health Research* 13 (4): 510–27. doi:10.1177/1049732302250721.

Atkins, Salla, Annika Launiala, Alexander Kagaha, and Helen Smith. 2012. 'Including Mixed Methods Research in Systematic Reviews: Examples from Qualitative Syntheses in TB and Malaria Control'. *BMC Medical Research Methodology* 12 (1): 62. doi:10.1186/1471-2288-12-62.

Bondas, Terese, and Elisabeth O. C. Hall. 2007. 'Challenges in Approaching Metasynthesis Research'. *Qualitative Health Research* 17 (1): 113–21. doi:10.1177/1049732306295879.

Bornemisza, Olga, M. Kent Ranson, Timothy M. Poletti, and Egbert Sondorp. 2010. 'Promoting Health Equity in Conflict-Affected Fragile States'. *Social Science & Medicine* 70 (1): 80–88. doi:10.1016/j.socscimed.2009.09.032.

Buhmann, Caecilie Böck. 2005. 'The Role of Health Professionals in Preventing and Mediating Conflict'. *Medicine, Conflict, and Survival* 21 (4): 299–311. doi:10.1080/13623690500268865.

Buhmann, Caecilie, Joanna Santa Barbara, Neil Arya, and Klaus Melf. 2010. 'The Roles of the Health Sector and Health Workers Before, during and after Violent Conflict'. *Medicine, Conflict and Survival* 26 (1): 4–23. doi:10.1080/13623690903553202.

Campbell, Rona, Pandora Pound, Catherine Pope, Nicky Britten, Roisin Pill, Myfanwy Morgan, and Jenny Donovan. 2003. 'Evaluating Meta-Ethnography: A Synthesis of Qualitative Research on Lay Experiences of Diabetes and Diabetes Care'. *Social Science & Medicine* 56 (4): 671–84. doi:10.1016/S0277-9536(02)00064-3.

Carroll, Christopher, Andrew Booth, and Myfanwy Lloyd-Jones. 2012. 'Should We Exclude Inadequately Reported Studies From Qualitative Systematic Reviews? An Evaluation of Sensitivity Analyses in Two Case Study Reviews'. *Qualitative Health Research* 22 (10): 1425–34. doi:10.1177/1049732312452937.

Dixon-Woods, Mary, Sheila Bonas, Andrew Booth, David R. Jones, Tina Miller, Alex J. Sutton, Rachel L. Shaw, Jonathan A. Smith, and Bridget Young. 2006. 'How Can Systematic Reviews Incorporate Qualitative Research? A Critical Perspective'. *Qualitative Research* 6 (1): 27–44. doi:10.1177/1468794106058867.

Dixon-Woods, Mary, Debbie Cavers, Shona Agarwal, Ellen Annandale, Antony Arthur, Janet Harvey, Ron Hsu, et al. 2006. 'Conducting a Critical Interpretive Synthesis of the Literature on Access to Healthcare by Vulnerable Groups'. *BMC Medical Research Methodology* 6 (1): 35. doi:10.1186/1471-2288-6-35.

Dixon-Woods, Mary, Alex Sutton, Rachel Shaw, Tina Miller, Jonathan Smith, Bridget Young, Sheila Bonas, Andrew Booth, and David Jones. 2007. 'Appraising Qualitative Research for Inclusion in Systematic Reviews: A Quantitative and Qualitative Comparison of Three Methods'. *Journal of Health Services Research & Policy* 12 (1): 42–47. doi:10.1258/135581907779497486.

Edwards, A. G., I. T. Russell, and N. C. Stott. 1998. 'Signal versus Noise in the Evidence Base for Medicine: An Alternative to Hierarchies of Evidence?' *Family Practice* 15 (4): 319–22. doi:10.1093/fampra/15.4.319.

Evans, Jennifer, and Pauline Benefield. 2001. 'Systematic Reviews of Educational Research: Does the Medical Model Fit?' *British Educational Research Journal* 27 (5): 527–41. aph.

Fischer, Arnout R H, Hilde Tobi, and Amber Ronteltap. 2011. 'When Natural Met Social: A Review of Collaboration between the Natural and Social Sciences'. *Interdisciplinary Science Reviews* 36 (4): 341–58. doi:10.1179/030801811X13160755918688.

Gutlove, Paula, and Gordon Thompson. 2006. 'Health As S Bridge For Peace: Achievements, Challenges, and Opportunities for Action by WHO'. Working Paper, World Health Organization, Department for Health Action in Crises.

Heyvaert, M., B. Maes, and P. Onghena. 2011. 'Mixed Methods Research Synthesis: Definition, Framework, and Potential'. *Quality & Quantity*, 1–18. doi:10.1007/s11135-011-9538-6.

Higginbottom, Gina M. A., Myfanwy Morgan, Jayantha Dassanayake, Helgi Eyford, Mirande Alexandre, Yvonne Chiu, Joan Forgeron, and Deb Kocay. 2012. 'Immigrant Women's Experiences of Maternity-Care Services in Canada: A Protocol for Systematic Review Using a Narrative Synthesis'. *Systematic Reviews* 1 (1): 27. doi:10.1186/2046-4053-1-27.

Hubbard, G., L. Kidd, and E. Donaghy. 2008. 'Involving People Affected by Cancer in Research: A Review of Literature'. *European Journal of Cancer Care* 17 (3): 233–44. doi:10.1111/j.1365-2354.2007.00842.x.

Hughes, Nic, S. José Closs, and David Clark. 2009. 'Experiencing Cancer in Old Age: A Qualitative Systematic Review'. *Qualitative Health Research* 19 (8): 1139–53. doi:10.1177/1049732309341715.

MacQueen, G, and J Santa-Barbara. 2000. 'Peace Building through Health Initiatives'. *BMJ (Clinical Research Ed.), 2000, Vol.321(7256), pp.293-6* 321 (7256): 293–96.

Magarey, Judith M. 2001. 'Elements of a Systematic Review'. *International Journal of Nursing Practice* 7 (6): 376–82. doi:10.1046/j.1440-172X.2001.00295.x.

Mahieu, Lieslot, and Chris Gastmans. 2012. 'Sexuality in Institutionalized Elderly Persons: A Systematic Review of Argument-Based Ethics Literature'. *International Psychogeriatrics* 24 (03): 346–57. doi:10.1017/S1041610211001542.

Major, Claire Howell, and Maggi Savin-Baden. 2012. *An Introduction to Qualitative Research Synthesis: Managing the Information Explosion in Social Science Research*. Routledge.

McCullough, Laurence B, John H Coverdale, and Frank A Chervenak. 2007. 'Constructing a Systematic Review for Argument-Based Clinical Ethics Literature: The Example of Concealed Medications'. *The Journal Of Medicine And Philosophy* 32 (1): 65–76. mnh.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. SAGE.

Noblit, G. W., and R. D. Hare. 1988. *Meta-Ethnography: Synthesizing Qualitative Studies*. Vol. 11. Sage Publications, Incorporated. http://books.google.com/books?hl=en&lr=&id=fQQb4FP4NSgC&oi=fnd&pg=PA10&dq=).+Meta-ethnography:+synthesizing+qualitative+studies.&ots=MT4Iwr81Rf&sig=1Ma0GNBD3wOlmo4uHix5YwY5oTI.

Paterson, Barbara L., Sally E. Thorne, Connie Canam, and Carol Jillings. 2001. *Meta-Study of Qualitative Health Research: A Practical Guide to Meta-Analysis and Meta-Synthesis*. SAGE.

Price, Eboni G. 2005. 'A Systematic Review of the Methodological Rigor of Studies Evaluating Cultural Competence Training of Health Professionals'. Edited by Mary Catherine Beach, Tiffany L Gary, Karen A Robinson, Aysegul Gozu, Ana Palacio, Carole Smarth, Mollie Jenckes, et al. *Academic Medicine : Journal of the Association of American Medical Colleges, 2005, Vol.80(6), pp.578-86* 80 (6): 578–86.

Roberts, Bayard, Eliaba Y. Damundu, Olivia Lomoro, and Egbert Sondorp. 2010. 'The Influence of Demographic Characteristics, Living Conditions, and Trauma Exposure on the Overall Health of a Conflict-Affected Population in Southern Sudan'. *BMC Public Health* 10 (1): 518. doi:10.1186/1471-2458-10-518.

Schaafsma, Frederieke, Jos Verbeek, Carel Hulshof, and Frank van Dijk. 2005. 'Caution Required When Relying on a Colleague's Advice; a Comparison between Professional Advice and Evidence from the Literature'. *BMC Health Services Research* 5 (1): 59. doi:10.1186/1472-6963-5-59.

Secomb, Jacinta. 2008. 'A Systematic Review of Peer Teaching and Learning in Clinical Education'. *Journal of Clinical Nursing* 17 (6): 703–16. doi:10.1111/j.1365-2702.2007.01954.x.

Sofaer, Neema, and Daniel Strech. 2011. 'Reasons Why Post-Trial Access to Trial Drugs Should, or Need Not Be Ensured to Research Participants: A Systematic Review'. *Public Health Ethics* 4 (2): 160–84. doi:10.1093/phe/phr013.

———. 2012. 'The Need for Systematic Reviews of Reasons'. *Bioethics* 26 (6): 315–28. doi:10.1111/j.1467-8519.2011.01858.x.

Wallace, A., K. Croucher, D. Quilgars, and S. Baldwin. 2004. 'Meeting the Challenge: Developing Systematic Reviewing in Social Policy'. *Policy & Politics* 32 (4): 455–70.

Weaver, Heather, and Bayard Roberts. 2010. 'Drinking and Displacement: A Systematic Review of the Influence of Forced Displacement on Harmful Alcohol Use'. *Substance Use & Misuse* 45 (13): 2340–55. doi:10.3109/10826081003793920.