Making a difference: towards a method for weighing the evidence in a qualitative synthesis

Hennie R. Boeije PhD,1 Floryt van Wesel MSc2 and Eva Alisic MSc3

1Associate Professor, 2PhD Candidate, Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, the Netherlands
3PhD Candidate, Psychotrauma Center for Children and Youth, University Medical Center Utrecht, Utrecht, the Netherlands

Keywords
evidence, qualitative research, qualitative synthesis, quality appraisal, systematic review

Correspondence
Dr Hennie R. Boeije
Department of Methodology and Statistics
Faculty of Social and Behavioural Sciences
Utrecht University
PO Box 80140
3508 TC Utrecht
The Netherlands
E-mail: h.boeije@uu.nl

Accepted for publication: 23 February 2011
doi:10.1111/j.1365-2753.2011.01674.x

Abstract

Objectives In a qualitative synthesis, primary qualitative studies are integrated to develop a theory or evidence-based interventions. Until now, the strength of the evidence in the primary studies has not been taken into account in the outcome of the qualitative synthesis. In this paper, a method is developed and evaluated to assign weights to the findings of the qualitative studies using both the frequency and the quality of the reported results.

Method Seventeen qualitative studies were retrieved in an illustrative example project on children and trauma. Findings were extracted from the primary studies with the use of coding which resulted in 14 themes. The quality of the studies was appraised using both expert judgement and a quality checklist. These outcomes are used to calculate the weights.

Results When the outcomes of the checklist appraisal are compared with those of the expert judgement, the effect on the strength of the evidence is virtually always in the same direction. We found that as the frequency with which a concept is studied is low, the strength of the evidence oftentimes decreases even further when using the quality of the results in the weighing process.

Conclusions In the end, the outcomes of a critical appraisal affect the weight that is placed on particular studies. The use of a checklist is recommended because of its more distinguishing ability. The method that was developed for assigning weights to the evidence is discussed in light of both the quality appraisal in qualitative research and the objectives of qualitative synthesis.

Introduction

As the number of qualitative studies in several scientific areas of interest increases, the need to systematically synthesize the findings, often referred to as qualitative synthesis or meta-synthesis, increases as well. The task of such a synthesis is ‘to bring together, juxtapose, re-analyse and combine the findings from several studies into a whole that ideally provides some theoretical or conceptual development that moves beyond the findings of any individual study included in the synthesis’ [1]. A qualitative synthesis has two broad objectives. The first objective is to gain a precise view of the theoretical knowledge in a focused area, as well as the possible knowledge gaps within that area. The second objective is to provide a knowledge base for the development of practice interventions, usually referred to as evidence-based interventions.

Synthesizing research evidence involves ascertaining how each individual study fits with the other studies in order to create the whole picture. One of the synthesis activities is assessing the methods and results of each study, often referred to as ‘critical appraisal’ or ‘assessing study quality’. In systematic reviews, assessing study quality is often used to refer to the extent to which the study lives up to methodological demands [2]. In other words, quality appraisal is concerned with methodological flaws that will change the ultimate outcomes of the synthesis. If quality appraisal is missing, it could potentially lead to over- or under-reliance on certain findings.

The appraisal of the quality of each individual study is done regularly in quantitative meta-analyses by differentiating clearly between higher and lower quality studies with the use of sample sizes [3]. In qualitative synthesis, few efforts have been made to use quality assessment to weight the studies to be included in the synthesis. Quality measurement is mainly used for screening for eligibility, that is, inclusion or exclusion. It is seldom used when looking at the adequacy of the evidence provided by the study. Here, quality appraisal appears to be a procedural step of little significance [4,5]. Above all, the absence of the weighting of
quality seems to be based on the assumption that all of the findings are of equal worth, which is a highly debatable assumption. We determined how many primary studies supported a certain theme and ordered the themes according to their frequency. The findings for each of the themes were described and an attempt was undertaken to synthesize them in an overall model. For the sake of clarity, we will use the themes – the clustered codes that constitute the building blocks of the findings – as our starting point in this article. The substantive contents of this research project on children and trauma will be described elsewhere (F. van Wesel, H.R. Boeije, E. Alisic & S. Drost, in preparation).

Methods

Search for articles

We chose a research on children and trauma as an illustration. The qualitative synthesis research project on children and trauma was carried out by the Department of Methodology and Statistics, Utrecht University and the Psychotrauma Center of the University Medical Center Utrecht in the Netherlands. The project focused on children under the age of 19 who have experienced some form of trauma. Trauma is defined as found in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and described as a sudden and unexpected life-threatening event or the witnessing of such an event in others that is experienced with anger, shock and helplessness [7].

The following electronic databases were searched: PsycINFO, PubMed, CINAHL, PILOTS and EMBASE. In addition, an issue-by-issue search was conducted of relevant scientific journals as well as hand searches through reference lists from literature reviews. Only peer-reviewed journal articles, written in English, and published between 1 January 1980 and 1 September 2009, were included in the search. Duplicates and studies that were not empirical reports and did not use qualitative methods or did not deal with trauma (from the children’s perspective) were removed. Of the remaining articles, 38 met the criteria for the children and trauma project. Six articles were untraceable. The other articles were read by the researchers in order to decide which articles should be included in the definitive sample for the synthesis. Fifteen articles were excluded for reasons such as trauma not being the main topic or the absence of qualitative methods. The final sample consisted of the 17 articles which are described in Table 1 [8–24]. To preserve anonymity, the articles are referred to with a randomly assigned letter.

Finding the findings

To extract the findings from the primary studies and synthesize them, we used the method of meta-data analysis [25]. The research findings of the studies were summarized by the authors and the summaries were discussed and compared for differences and commonalities. The first two authors analysed the summaries of five studies by coding them for themes [26] using Maxqda2007 (Verbi, Marburg, Germany), computer software for qualitative analysis. The codes that were assigned to the summaries were compared and discussed, resulting in one code system. The summaries of all of the remaining studies were coded by the second author and few new codes that were generated in this step were discussed with the others and agreed upon.

The separate codes in the code system were clustered, which resulted in 14 themes that could be grouped in an overall structure consisting of three domains: individual, family and community.
combined by article weights referring to the quality appraisal with the quality checklist. The weights were determined using the following five-step procedure:

Step 1: Calculate an average appraisal score per article \((a = 1, \ldots, A)\) per appraisal method (expert judgement or checklist) by taking the mean score of the two or three reviewers (\(d_a\)). These mean appraisal scores (for the expert judgement on a 1-to-10 scale and for the checklist rescaled from a 0-to-20 scale – as there were 10 items and two possible points per item – to a 1-to-10 scale) can be found in the fifth and ninth column of Table 2.

Step 2: Calculate the mean score over all articles (\(m \equiv \sum d_a/k\), where \(k\) denotes the total number of included articles per appraisal measure (see final row of Table 2) on a 1-to-10 scale (rescaled scores for the checklist). As a result, we get one mean score for the checklist (\(m_c\)) and one mean score for the expert judgement (\(m_e\)) averaged over all reviewers and articles.

Step 3: For each article, a code whether or not (0 or 1) a certain theme \((b = 1, \ldots, B)\) is present (\(c_{ba}\)).

Step 4: Calculate the evidence weight per article per appraisal method:

\[
weight_a = \left(\frac{d_a - m}{2}\right) + 1,
\]

where \(weight_a\) denotes the weight per article based on the checklist or the expert judgement. The difference scores are divided by 2 to prevent the weights from being negative (the largest difference between any \(d_a\) and \(m\) is 2).

Step 5: Calculate amount of evidence per theme \(b\) based on all articles:

\[
evidence_b = \sum c_{ba} \cdot weight_a,
\]

where \(evidence_b\) is the amount of evidence for theme \(b\).

---

**Table 1 Characteristics of the primary articles in the qualitative synthesis**

<table>
<thead>
<tr>
<th>No.</th>
<th>Reference</th>
<th>Type of trauma</th>
<th>Sample</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Al-Mashat et al. [8]</td>
<td>War</td>
<td>12 children</td>
<td>Data: focus groups, drawings</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 9–13</td>
<td>Analysis: focus group methods</td>
</tr>
<tr>
<td>2.</td>
<td>Coholic et al. [9]</td>
<td>Chronic trauma</td>
<td>38 children</td>
<td>Data: group session, interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20 foster parents</td>
<td>Analysis: constant comparison methods</td>
</tr>
<tr>
<td>3.</td>
<td>DeVoe &amp; Smith [10]</td>
<td>Domestic violence</td>
<td>43 mothers</td>
<td>Data: focus groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 2–6</td>
<td>Analysis: modification of coding system</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 1–14</td>
<td>Analysis: thematic analysis</td>
</tr>
<tr>
<td>5.</td>
<td>Forinder &amp; Lindahl Norberg [12]</td>
<td>Brain tumour</td>
<td>11 parents</td>
<td>Data: open-ended interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 7–14</td>
<td>Analysis: inductive thematic method</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 13–17</td>
<td>Analysis: constant comparison methods</td>
</tr>
<tr>
<td>7.</td>
<td>Horowitz et al. [14]</td>
<td>Community violence</td>
<td>28 children</td>
<td>Data: focus groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23 parents</td>
<td>Analysis: thematic coding</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 8–17</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 13–15</td>
<td>Analysis: grounded theory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 17–18</td>
<td>Analysis: unclear</td>
</tr>
<tr>
<td>10.</td>
<td>Moscardino et al. [17]</td>
<td>Terrorism</td>
<td>17 caregivers</td>
<td>Data: participant observation, semi-structured interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 6–14</td>
<td>Analysis: thematic content analysis, code development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 11–17</td>
<td>Analysis: topical recording, keyword recording</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 15–17</td>
<td>Analysis: interpretative phenomenological analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17 parents</td>
<td>Analysis: unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 6–16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 7–18</td>
<td>Analysis: framework technique</td>
</tr>
<tr>
<td>15.</td>
<td>Thastum et al. [22]</td>
<td>Parental cancer</td>
<td>21 children</td>
<td>Data: semi-structured interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 8–15</td>
<td>Analysis: descriptive phenomenological method</td>
</tr>
<tr>
<td>16.</td>
<td>Urman et al. [23]</td>
<td>Traumatic experiences</td>
<td>6 children</td>
<td>Data: semi-structured interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 9–13</td>
<td>Analysis: grounded theory</td>
</tr>
<tr>
<td>17.</td>
<td>Woodgate &amp; Degner [24]</td>
<td>Childhood cancer</td>
<td>39 children</td>
<td>Data: focus groups, interviews, reflexive journals, participant observation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age: 4–18</td>
<td>Analysis: constant comparative method, illness narrative inquiry</td>
</tr>
</tbody>
</table>

© 2011 Blackwell Publishing Ltd
When a weight is smaller than 1, it means that the quality of an article is lower than average (given the appraisal method), and when a weight is larger than 1, it means that the quality of an article is better than average. As a result of calculating the weight in this way, all three (weighted) evidence outcomes – frequency, frequency times expert judgement and frequency times checklist – are measured on the same scale and can therefore be easily compared.

Results

Appraisal with expert judgement

The scores that the articles received from the different reviewers are presented in Table 2. The reviewers assigned, on average, 6.8 for quality with the expert judgements. The lowest mark was 5.0 and the highest 8.3. Typical remarks about articles with a low score (<5.5) were: ‘I find the presentation chaotic and unsystematic’ (reviewer 1). A typical comment on an article with a high score (>8.0) was: ‘The analysis is clear and accurate. The interpretations are discussed with the research team members and they obviously are aware of any limitations of their research’ (reviewer 2).

There was a high inter-rater reliability between the reviewers (intra-class correlation = 0.88). However, the differences in the expert judgements were sometimes quite pronounced. Five articles had a difference in scores of 1.5. In the discussion between the researchers about the outcomes based on our expert judgements, the claims stood out that were made in the studies, given their research questions and the studies’ objectives. We argue that studies that are intended to provide an inventory and a description of themes, sometimes referred to as thematic surveys [30], were troubled with under-analysis; when looking at the available data, we concluded that more could have been done with it. Studies that are intended to provide a fully integrated explanation of a phenomenon, event or case, as in grounded theory or ethnography, were sometimes found to have an over-analysis, producing, for instance, an entire grounded theory on the basis of only a few interviews. We felt that they overplayed their hand, but this did not lead, in general, to a low score.

Appraisal with the checklist

On average, the articles received 6.9 for quality with the use of the checklist by all reviewers (Table 2). The lowest score was 4.7 and the highest 9.3. Again, there was a high inter-rater reliability between the reviewers (intra-class correlation = 0.94). The differences in the checklist appraisal were less pronounced than in the expert judgement.

Items 7, 8 and 10 of the checklist scored relatively low (<6.0) in most of the articles. Item 7 received low scores, because in most of the studies the authors neither elaborated on specific quality issues, such as triangulation, member checks, discussion between team members, nor paid attention to the limits and strengths of their studies. Item 8 scored low, because several studies did not speak about any ethical issues. Item 10 scored low, because in most primary studies no attention was paid to descriptions of the sample or to generalization options.

The items 1, 3 and 9 scored relatively well (>8.0). Item 1 scored well, because most studies had quite elaborate literature reviews and posed relatively good research objectives. Item 3 also received quite high values, because most authors provided detailed accounts of data collection methods. Item 9 received good scores, because most reports were well structured and used conventional reporting styles.
Comparison between expert judgement and checklist

The correlation between the overall expert judgement and the checklist appraisal was 0.867, which is considered quite high. There are three articles, [D], [H] and [P], with more than 1-point difference in scores when the checklist is compared with the expert judgement. Article [H] has high scores on the checklist on almost all items. However, the expert judgement was low. According to reviewer 1, this has to do with the confusion about the subject matter that the article deals with and with the lack of internal consistency. The scores of article [D] are the opposite to those of article [H]. With the checklist, it scores quite low, but regarding the significance of the findings and coherence, the reviewers rate it favourably. Article [A] shows a large difference between both the reviewers and the methods. One of the reviewers thinks the article is quite good, scoring moderately high with the checklist and the expert judgement, while the other reviewer thinks it is weak on one aspect, in particular the sources of the findings being either parents or children, and this appraisal contaminates the entire article.

Weighting the evidence

The outcomes of the appraisals are now used to weigh the evidence that each of the primary articles provides for the findings, here the themes or building blocks. In this section, evidence is weighted in three ways. First, we only look at the number of times that a certain theme appears in the primary studies, the ‘frequency’. Second, we combine this frequency and the expert appraisal. Third, the frequency is combined with the quality appraisal using the checklist (Fig. 1).

The 14 themes that resulted from the analysis of the primary studies and the three overall domains are presented on the x-axis of Fig. 1. The amount of evidence found for each concept is represented by the height of the bars on the y-axis. The concepts are ordered by their frequency, per domain.

The height of the black bars is determined by the frequency with which each concept is found in the primary articles (Note that this is an across-reports frequency and not a within-report frequency). A theme is either present or absent in a primary article. Four themes are found in one to six primary studies. Seven themes are found in 7 to 11 articles. Finally, the three remaining themes are supported by 11 or more studies.

The height of the grey bars is determined by the frequency combined with the scores of the expert appraisal. When weight is calculated using the frequency and the expert appraisal, the height of the bars of a specific concept changes. The evidence for the concepts of feelings, trauma impact, coping, giving meaning, current outlook, phases, parenting and interpersonal relationships increases considerably. The amount of support for the concepts of negotiation and normalcy increases only slightly. The support for the concepts of growth, support and culture decreases and the evidence for identity decreases only slightly.

The height of the white bars represents the combination of the frequency combined with the checklist appraisal. Themes with considerably more support when quality is weighed are feelings, trauma impact, coping, giving meaning, current outlook, phases and parenting. Evidence in favour of the themes, negotiation and interpersonal relationships, increases slightly. The strength of the evidence for identity, normalcy, growth, support and culture declines.

When the methods for weighting are compared, we noticed that the direction of the change of a concept as a consequence of the introduction of quality is always in the same direction (higher or lower for both quality appraisals) except for the concept normalcy. This theme moves upwards with the expert judgement and moves downwards with the checklist appraisal. When evidence is weighted with the checklist appraisal, the changes are, in most cases, consistent with the expert judgement.
cases, more pronounced than with the weights of the expert judgement. This can be seen in Fig. 1, when comparing the grey and white bars. On average, the evidence weighted by the checklist differs more from evidence based on frequency only, than the evidence weighted by the expert judgements. This leads us to conclude that the checklist has a somewhat higher differentiating ability.

As can be seen in Fig. 1, the order of the themes organized by amount of evidence would change because of the weighing with the quality appraisal. Although identity scored higher than current outlook and phases, when only frequency is used as evidence, when the quality appraisal with the checklist is introduced, the evidence for the concepts of current outlook and phases gets more support than identity. In other words, identity is more often examined than current outlook and phases, but in studies which are of relatively less (methodological) quality. In general, themes that decline because of the weighing with the quality appraisal (normalcy, growth, culture) are already less frequently studied.

**Discussion**

In this article, an attempt is made to weight the evidence for the findings of the primary studies in a qualitative synthesis project. The frequency as well as the quality of the primary studies was used to weight the strength of the evidence. The line of argument here is that when a topic is frequently studied in a methodologically sound way, there is strong evidence for the value of the findings from a methodological point of view. In contrast, when a topic is seldom studied and the studies are methodologically weak, the evidence for the findings is also weak. In (quantitative) meta-analysis, sample sizes of the original studies define the weights of their effect sizes when they are combined, because effect size values based on larger samples are more precise. In addition, some meta-analysts include methodological quality in the weighing procedure, for example, taking into account random sampling procedures [3]. Such a weighting of evidence has been absent in qualitative research synthesis efforts.

Weighting the evidence in qualitative synthesis is recommended as a valuable tool for correctly estimating the value of the evidence. This is necessary to prevent mistakes in the outcomes of qualitative syntheses whether in theory building or in the development of practical interventions. As was clear from the results, some of the themes in the qualitative synthesis outcomes received more support and others received less support after weighing the quality of the primary studies. Not all studies are of the same quality level and this should be accounted for in integrative efforts. Edwards et al. [6] proposed to submit a collection of studies on a particular subject to ‘signal’ and ‘noise’. A publication (study) is first examined for its relevance, applicability and value (signal) and then for its strengths and weaknesses (noise). With our research, we have contributed to examining noise and not to assessing the substantive findings of the primary studies (signal).

We believe our procedure to be in line with the suggestion of Dixon-Woods et al. [4] to consider all studies (and not to exclude any of them beforehand on the basis of a quality judgement) and to then build in an appraisal of methodological limitations of the studies. This so-called critique-led approach gives weight to the argument that qualitative research deals mainly with the conceptual contribution and that, as a consequence, all input is welcomed [33]. The outcomes of this method could be compared with the use of a sensitivity analysis, which involves testing how sensitive the review findings are to the inclusion and exclusion of studies of different quality. The question to be answered is: what would happen to the results, if all the studies below a certain established ‘quality threshold’ would systematically be excluded?

When the outcomes of the checklist appraisal are compared with the outcomes of the expert judgement, the effect on the strength of the evidence is virtually always in the same direction. Either the evidence for the concept, that is, the clustered findings of the primary studies, increases or it decreases. The outcomes of the checklist appraisal are somewhat more pronounced. We found that the strength of the evidence diminishes, especially for the concepts that are already less frequently studied. On the one hand, this means that they are not only less frequently studied, but that they appear in studies that are methodologically not quite adequate. On the other hand, this is considered to be an indication that these concepts need further research that is methodologically sound.

In our attempt towards generating a method for weighting the evidence of findings in a qualitative synthesis, a formulaic approach was used. Although we believe to have demonstrated that such an approach is feasible, several aspects need further reflection. First, working with a formula is possible when the themes or building blocks of the findings of the primary studies are being used, but this could work to the detriment of the complexity of and the detailed nature of the findings themselves. Second, we used a summary score of the individual items in the checklist. Petticrew and Roberts [2] warn that a summary score hides the justifiable judgements and trade-offs within a study and conceals errors that could be considered fatal and that could eventually lead to the exclusion of a study. Third, a formula was used that included the values of the critical appraisal and expert judgement in combination with the frequency. Further research is needed to examine the validity and reliability of this weighting scale.

Considering the more highly discerning ability, a checklist appraisal is being recommended over an expert judgement. An additional reason for this recommendation is that the checklist appraisal is transparent for team members as well as for others and can easily be discussed. Two conditions will optimize the use of a checklist. The first is that different reviewers need to be involved and discuss their appraisals in order to realize a high consensus on the items and the scores assigned to the primary studies. The second condition is that the checklist needs to cover the opinions on the items by the researchers. If necessary, this can be accomplished by adding an additional item that is similar to the expert judgement. We feel that this will encourage researchers to actually use the outcomes of the quality appraisal in weighing the evidence. Literature suggests that it is a hindrance to use the outcomes of a checklist appraisal when researchers feel that these do not match with their own judgement [31].

Finally, we all felt that discussing our quality appraisals in an interdisciplinary team had a positive and uplifting effect on the entire synthesis project, despite its time-consuming and labour-intensive character. We felt that forming a judgement as part of a team sharpened our minds and encouraged us to put our views into words. It became very clear to us that a very valuable aspect of quality appraisal is that it makes all team members aware of all the different types of studies. We firmly believe that this will ultimately lead to more carefully constructed findings.
References


Appendix

Checklist used for quality appraisal of primary studies

1. Research purposes or questions are linked to the problem and/or to the review of the literature.

2. Sample size and configuration fit the purpose and sampling strategy.

3. Data collection techniques are correctly used.

4. Data analysis techniques are tailored to the reported study.

5. Interpretations of data are demonstrably plausible and/or sufficiently substantiated with data.

6. The clinical, policy, theoretical, disciplinary, and/or other significance of the findings is thoughtfully considered.

7. The study discusses techniques specifically intended to ensure that the study is scientifically and/or ethnographically valid or ‘good’.

8. Benefits and risks distinctive to the study are addressed.

9. Given the reporting style, elements of the research report are placed where readers are likely to find them.

10. The authors provide information regarding participants, setting and context so that the reader might be able to determine the relevance of the findings to other settings (transferability).