Qual Quant (2015) 49:2237–2249
DOI 10.1007/s11135-014-0105-9

Towards a method for synthesizing diverse evidence using hypotheses as common language

F. van Wesel · H. R. Boeije · E. Alisic

Published online: 15 October 2014

Abstract Combining the findings obtained by different research methods in mixed-research synthesis could potentially contribute to a broader, more diverse evidence base for interventions. In this article we focus on the methodological challenges involved in synthesizing various types of research findings. We propose a method that uses hypotheses to facilitate the comparison and integration of such different findings. The method consists of four steps: (1) synthesizing findings per source of evidence, (2) formulating a mono-method hypothesis for each source, (3) integrating the monomethod hypotheses into one overall hypothesis, and (4) evaluating, using empirical data, whether the overall hypothesis better fits the data than each of the mono-method hypotheses. Using quantitative studies, qualitative studies and experts' views in the substantive case of children and trauma, we will illustrate the proposed method. We conclude that the method provides a viable perspective for constructing an elaborate model that captures the knowledge from complementary sources.

Keywords Directional hypotheses · Evidence synthesis · Meta-analysis · Mixed-research synthesis · Qualitative synthesis

1 Introduction

The aim of the evidence-based movement which began in the mid-1970s was to practice medicine based on a systematic and thorough investigation of the current evidence. Although at first the results of solely randomized controlled trials (RCTs) were used to answer questions

F. van Wesel (✉)
Department of Educational Neuroscience & Department of Methodology, Faculty of Psychology and Education, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
e-mail: f.van.wesel@vu.nl

H. R. Boeije
Department of Methodology and Statistics, Utrecht University, Padualaan 14 Room C.115, 3584 CH Utrecht, The Netherlands

E. Alisic
Monash Injury Research Institute, Monash University, Building 70, Melbourne, VIC 3800, Australia

concerning the effectiveness of treatments, in the past decade other questions relating to policy making and professional practice arose (Dixon-Woods et al. 2006a). These questions pertained, for example, to the acceptability of programs for certain target groups, to professional competences and preferences in program implementation, to working mechanisms that explain the (in-) effectiveness of interventions, and to cost-effectiveness (e.g., Dixon-Woods et al. 2005; Pope et al. 2007). It was acknowledged that the question 'What intervention works best' which matched the design of the RCT could be expanded with other questions such as 'Why does it work' and 'Under what circumstances?'. These questions match studies using other methodologies such as qualitative research or expert elicitation (e.g., Barbour 2000; Dixon-Woods et al. 2001).

Studies that combined the outcomes of quantitative and qualitative research were successful in answering review questions that applied to complex issues in policy-making and practice. Harden and Thomas (2005), for example, examined the promotion of healthy food for children. They combined the results of meta-analyses that measured effectiveness with the results of a qualitative synthesis that determined children's preferences (Thomas et al. 2004). Other examples using this method examined therapy adherence (Candy et al. 2011), teenage pregnancy in socially-disadvantaged groups, and other life-style issues (Harden et al. 2009). A different method for combining qualitative and quantitative research, Bayesian meta-analysis, was used for studying immunization uptake (Roberts et al. 2002). For more examples, Gorecki et al. (2009) and Sandelowski et al. (2007). In addition to the methods mentioned above, several other methods were proposed (e.g., Dixon-Woods et al. 2006a; Pope et al. 2007; Pawson et al. 2004).

The outcomes of quantitative and qualitative research are difficult to compare because they use numerical and textual data, respectively (Sandelowski et al. 2006). In addition, they differ with respect to design, sampling, theorizing, analysis, and reporting. In their overview, Dixon-Woods et al. (2005) outline several methods for mixing both strands of research on a meta-level. With a few exceptions, most methods described here were originally developed to process only one strand. Although most methods could be modified to accommodate both sources, this modification would mainly rely on transforming either qualitative data into numerical data or quantitative data into textual data. Consequently, the (transformed) data would be analyzed with, respectively, statistical or qualitative analysis techniques. What can be concluded from this overview study is that incorporating qualitative research in the evidence base, without quantifying the results, remains methodologically challenging.

In this paper we propose an alternative method that can accommodate different sources of complex evidence. The method has ties with that of Harden and Thomas (2005) but can also handle evidence other than interventions. The core of the method we propose is the generation of a hypothesis on the basis of each source of evidence. These hypotheses then are comparable. Hypotheses can be formulated in words as well as in a statistical expression (formula). With this method the findings of each source included in the integration, are reduced to a single hypothesis. Hypotheses can contain a considerable amount of information within one single statement, which makes the method simultaneously complex yet simple. We will illustrate the method with a worked example on determinants for the development of post-traumatic stress disorder (PTSD) in traumatized children. Although we use qualitative research, quantitative research and an elicitation study with experts in this example, we believe the method to be appropriate for all sources that allow hypotheses to be formulated on the basis of their findings. The development and application of the proposed method will be shown in the following section.

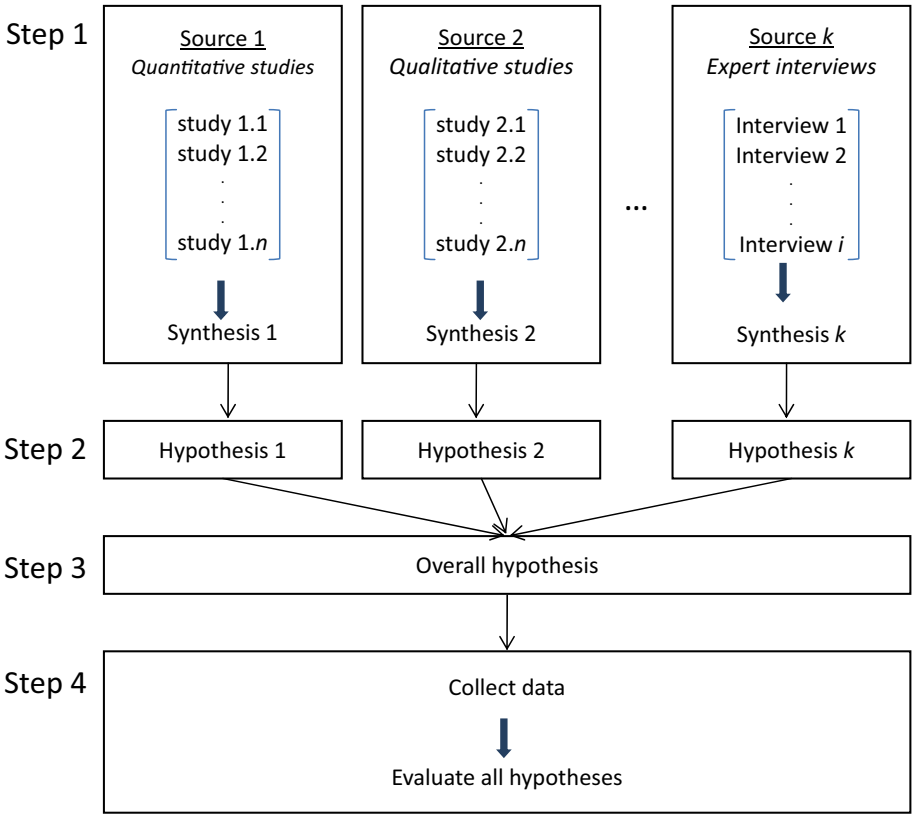

Fig. 1 Graphical representation of steps in the evidence synthesis using $k = 1, \ldots, K$ hypotheses

2 Evidence synthesis method

The core element of the method described in this study is the hypothesis. A hypothesis can be stated using words or statistical symbols. In this study we use the latter, i.e., hypotheses as statistical formulas. The outcomes of the sources, all using different languages, will be translated into hypotheses and expressed in formulas. This enables us to integrate the different sources of evidence into one overall result which will also be stated as a hypothesis. The origins of the hypotheses, how they are formulated, how they are integrated and how they are evaluated is graphically represented in Fig. 1.

First, different sources are considered (Step 1). In Fig. 1 a source may represent a single study or a set of studies using a similar methodology (e.g., quantitative studies, qualitative studies, policy documents). In our proposed method the number of sources is not fixed and is therefore represented as source 1–k in Fig. 1.

For reasons of simplicity, each source is used to generate one hypothesis. Note however that in practice one source may lead to several stand-alone hypotheses. Since this hypothesis is based on a one-method study(-set) it is referred to as a mono-method hypothesis (Step 2). These mono-method hypotheses are then synthesized to constitute one overall hypothesis (Step 3). Finally, this integrated hypothesis is tested on empirical data (Step 4). We will now continue with a more detailed description of the steps taken in the proposed method.

2.1 Step 1: synthesizing findings per source

We start with integrating the findings per separate source. Each source has its own specific methodological characteristics, e.g., numerical or textual data, which makes it possible to compare and integrate the findings. Therefore, the first step consists of synthesizing findings within a single source, for all k sources. The main advantage of this step is that for synthesizing findings from a single source, existing valid methods can be employed, such as meta-analysis for quantitative studies (Lipsey and Wilson 2001), meta-ethnography for qualitative studies (Noblit and Hare 1988), and qualitative analysis using constant comparison for interview data (Boeije 2009). Consequently, synthesizing single source findings requires no new methodology or requirements to determine the quality standards set for these kinds of research endeavors.

A second advantage of beginning with the integration of mono-method studies is that existing synthesis studies can be used. The number of synthesis studies is increasing each year and the number of qualitative syntheses is catching up with the number of quantitative reviews (Hannes and Macaitis 2012). These 'ready-to-go' sources contribute to the efficiency of the method proposed in this paper. Saving time by using ready-to-go sources might even lead researchers to include an extra source for which they themselves would have to conduct the synthesis or, to collect empirical data. This enlarges the evidence base resulting from the method, as even more sources are incorporated.

A third strength of synthesizing the findings per source as a first step is that the outcomes of the separate sources can easily be compared and their differences possibly explained by differences in the methodology used. For example, cross-sectional quantitative studies may reveal that age is an important determinant for developing PTSD, whereas longitudinal studies may point out that self-esteem is important. Findings such as these might lead researchers to focus on the effects of the outcomes resulting from methodological differences. This can enlarge the researchers' insights into the differing natures of the sources as well as into how the sources complement one another.

2.2 Step 2: formulating mono-method hypotheses

The second step consists of formulating separate hypotheses for each source outcome resulting from Step 1. If k sources are used this will result in k hypotheses. This process lies at the heart of the method for evidence synthesis that we present here. For our purposes a particular kind of hypothesis is chosen, referred to as a directional hypothesis, also known as an 'inequality constrained hypothesis' (Hoijtink et al. 2008). These hypotheses contain information on (a) the variables of interest and (b) relationships between the variables, expressed in order restrictions. When, for example, we are interested in finding the determinants of the development of PTSD, the variables of interest are predictors and they can be ordered according to their level of prediction.

In order to establish a directional hypothesis an appropriate statistical model needs to be chosen, in this example a multiple regression model. A hypothesis within this context might look like this: H: $\beta_a > \{\beta_b, \beta_c\} > \beta_d$, where β indicates a standardized regression coefficient, the subscript denotes the variable of interest, '$>$' denotes larger than (also possible are $<$ and $=$), ',' denotes no constraint between two parameters and $\{\}$ denotes a subset. At the end of this step, a mono-method hypothesis is formulated for each source. (From this point on, the terms determinant and predictor are used interchangeably.)

2.3 Step 3: integrating mono-method hypotheses

The third step consists of the bringing together of the mono-method hypotheses. The determinants that have been found in the different sources need to be compared. This comparison addresses differences and similarities in the determinants found, e.g., the first source shows that age is an important predictor for PTSD but this predictor was not revealed in the second source. In addition, the comparison is also directed at the differences in predictive value of each of the determinants. Based on the outcomes of the comparison, one hypothesis will be formulated.

In integrating the information of the different sources, it is appropriate to weigh the obtained information. We consider three possible methods for weighing. The first method is weighing for reliability of the findings based on the number of studies. A determinant resulting from a mono-method hypothesis based on 30 studies will receive more weight than a determinant resulting from a mono-method hypothesis based on 5 studies. The second method is weighing for the reliability of the findings based on the frequency of sources. In that case a predictor present in one source will receive less weight than a predictor present in all sources. The third method is weighing for the effect size of a determinant averaged over all sources in which it is present. In this case the overall hypothesis results from weighted information on the reliability of the determinants and their (averaged) effect size.

2.4 Step 4: evaluating all hypotheses

The fourth and final step of our proposed method is the evaluation of the formulated hypotheses. In this step we seek an answer to the question 'What are the important determinants and what is their predictive value?'. In addition to answering this question we also want to examine whether combining the different sources has surplus value over the mono-method hypotheses. Subsequently, we evaluate which of the hypotheses is the most likely, either one of the mono-method hypotheses or the overall hypothesis. This evaluation involves collecting new empirical data on all determinants that have been found in the mono-method syntheses. These data then will be analyzed using a Bayesian model selection procedure for inequality constrained hypotheses (Hoijtink et al. 2008; Mulder et al. 2009; van Wesel et al. 2010).

3 Worked example: children and trauma

Our example of an evidence synthesis is driven by the following review question: what are the risk factors and protective factors, i.e., the determinants, of the development of PTSD in children who have gone through trauma and how important are these factors? The sources that we use are a meta-analysis, a qualitative synthesis and the views of experts (collected in interviews). Both the meta-analysis (Alisic et al. 2011) and the qualitative synthesis (van Wesel et al. 2011) have been published elsewhere and details about the effect sizes can be found in these publications. The methods used for collecting and analyzing the expert views can be found in Appendix.

3.1 Step 1: synthesizing findings per source

The first source we use is that of quantitative studies, combined using meta-analysis. A meta-analysis is a method that evaluates the results of several conceptually-equal quantitative studies by means of effect sizes. The effect sizes will be combined into one statistic, with the

possibility to weigh for (un)reliability via sample size (e.g., Hunter et al. 1982; Lipsey and Wilson 2001).

The meta-analysis used in this synthesis endeavor is reported in Alisic et al. (2011) and involves prospective studies. The aim of this meta-analysis was to find determinants for the development over time of PTSD in traumatized children. The effect size chosen was the product moment correlation coefficient. Thirty-four studies were included in the analysis. The analysis resulted in statistics calculated for 12 variables. Three of these predictors (age, ethnicity, and social economic status) were found to be non-significant. The nine significant predictors and their effect sizes can be found in the second column of Table 1.

We use a qualitative synthesis as our second source. The aim of a qualitative synthesis is to produce a new and integrative interpretation of findings that is more substantive than an interpretation resulting from individual investigations (Sherwood 1999). The results of the single studies are coded, interpreted and integrated (e.g., Noblit and Hare 1988; Paterson et al. 2001; Pope et al. 2007; Sandelowski and Barroso 2007).

The focus of the qualitative synthesis (van Wesel et al. 2011) used here to formulate the second hypothesis was to investigate the trauma experience of children and how they worked their way through these experiences. Seventeen articles were included. The analysis of the included studies was done using meta-data-analysis (Paterson et al. 2001) and resulted in several qualitative themes. For the current example we identified those themes that had predictive value in PTSD development. The third column of Table 1 show the found effect sizes.

The third source used in our example are Dutch experts within the field of children and trauma who were interviewed about their perspectives on the topic. The aim of the interviews was to allow the experts to express their ideas about the determinants of the development of PTSD in children. We were able to get a purposive sample of six experts. The interviews were analyzed according to the principles of qualitative analysis (e.g., Boeije 2009). The determinants and their effect sizes, can be found in the fourth column of Table 1.

3.2 Step 2: formulating mono-method hypotheses

For the meta-analysis we were able to formulate a hypothesis that fits the multiple regression context, since the correlation coefficients used in the meta-analysis are closely related to regression coefficients. Note, however, that we shift from a univariate context in the meta-analysis to a multivariate context for the purpose of the research, which means that weak correlations may vanish in the presence of stronger correlations. To formulate a directional hypothesis, the weighted mean effect sizes are categorized using the rules of thumb proposed by Cohen (1992; the weighted mean effect sizes can be found in the final column of Table 1). This process leads to the directional hypothesis (H_{MA}):

$$\text{Hypothesis meta-analysis}: \{\beta_{\text{Anxiety}}, \beta_{\text{Depression}}, \beta_{\text{Acute stress}}, \beta_{\text{PTS (1, 3) months}}\} > \{\beta_{\text{Heart rate}}, \beta_{\text{Days in hospital}}, \beta_{\text{Parents' PTS}}\} > \{\beta_{\text{Gender}}, \beta_{\text{Injury severity}}\}.$$

To generate a hypothesis based on the qualitative synthesis, it was necessary to determine the importance of each determinant found in the qualitative synthesis. For this purpose the number of articles in which the determinant was present was counted. In formulating a hypothesis, the predictors were assigned an 'effect size'. In assigning effect sizes we aimed at having 1/3 of the themes in each of the three effect size categories. Therefore, determinants found in 10–17 articles were considered to be of high importance and were assigned a

Table 1 Determinants per source (weighted mean or frequency) and their effect size (S = small, M = medium, L = large) per source, the mean effect size over the three sources (label and numerical) and the ordering of determinants using two weightings [W_1, W_2, where $\max(a) = 3$ and $\max(b) = 5$]

	Meta-analysis		Qualitative synthesis		Experts	Mean ES		W_1	W_2
Feelings	0.45	L	14	L	L	L	5	15	15
Parenting	0.34	M	11	L	L	L/M	4	14	12
Injury severity	0.09	S			M	M/S	2	7	5
Acute reactions	0.51	L			S	M	3	8	8
Child characteristics			9	M	M	M	3	8	8
Coping			10	L	M	L/M	4	9	11
Support			7	M	M	M	3	8	8
Interpersonal relations			9	M	S	M/S	2	7	5
Culture			5	S	S	S	1	6	2
Gender	0.13	S				S	1	1	1
Heart rate	0.18	M				M	3	3	7
Days at hospital	0.18	M				M	3	3	7
PTS (1, 3) months	0.56	L				L	5	5	13
Trauma impact			12	L		L	5	5	13
Normalcy			5	S		S	1	1	1
Current outlook			7	M		M	3	3	7
Type of trauma					L	L	5	5	13
Ordering chaos					L	L	5	5	13
Safety					L	L	5	5	13
Trust					S	S	1	1	1

large effect size, determinants found in 7–9 articles were assigned a medium effect size and determinants present in 1–6 articles were considered to have a small effect size. The assigned effect sizes can be found in Table 1. This operation resulted in the following directional hypothesis (H_{QS}):

$$\text{Hypothesis qualitative synthesis: } \{\beta_{\text{Feelings}}, \beta_{\text{Trauma impact}}, \beta_{\text{Parenting}}, \beta_{\text{Coping}}\} > \{\beta_{\text{Identity}}, \beta_{\text{Interpersonal relationships}}, \beta_{\text{Current outlook}}, \beta_{\text{Support}}\} > \{\beta_{\text{Normalcy}}, \beta_{\text{Culture}}\}.$$

Finally, we wanted to integrate the experts' views while at the same time maintaining their specific foci. We established which predictors were mentioned most frequently by the experts and combined the result with a weight of each expert's idea about the strength of the predictor. In this way we were able to rank the predictors in a similar manner as we did in the qualitative synthesis. This resulted in the following hypothesis (H_{EV}):

$$\text{Hypothesis experts: } \{\beta_{\text{Type of trauma}}, \beta_{\text{Ordering chaos}}, \beta_{\text{Feelings}}, \beta_{\text{Safety}}, \beta_{\text{Parenting}}\} > \{\beta_{\text{Severity}}, \beta_{\text{Avoidance}}, \beta_{\text{Child characteristics}}, \beta_{\text{Support}}\} > \{\beta_{\text{Trust}}, \beta_{\text{Care after trauma}}, \beta_{\text{Culture}}, \beta_{\text{Interpersonal relationships}}\}.$$

3.3 Step 3: integrating mono-method hypotheses

In this step the separate hypotheses are compared, looking at the predictors that they include as well as at the difference in predictive value, i.e., effect size, of each of the predictors (see Table 1). When comparing the determinants of the separate mono-method hypotheses it becomes apparent that some determinants are present in all three sources, for example, *Feelings* and *Parenting*. This might indicate that these predictors are of great importance in the development of PTSD. However, the operational definition of such a predictor might differ per source. For example, the determinant *Parenting* (parents who are present during the traumatic event are a protective factor; absent parents are a risk factor) shares the definition used by the experts and also used in the qualitative synthesis (parenting represents an element of the parent–child interaction and refers to raising a traumatized child). In the meta-analysis however, this determinant is represented by the variable *parents' PTS*, which measures only parental stress reactions. Partly because of such differences the predictors that occurred in more than one source had varying predictive values (effect size) in the sources. For instance, *Parenting* is a medium predictor in the meta-analysis but a large predictor in the other two sources. Such contradicting results can be averaged out (see Sect. 3.3.1 below) or separate contradicting hypotheses can be formulated for evaluation in the final stage.

In contrast to these shared determinants, source unique determinants can be found. Such determinants could well be a product of the method used for data collection and analysis. For example, in the meta-analysis we find typical countable predictors that are typically countable such as the number of days a child had to stay in the hospital. In the qualitative synthesis we find a theme called 'Normalcy' representing a basic-psychosocial process, i.e., the continuous comparison between the pre- and the post-trauma world, which can be considered as a typical qualitative focus. When formulating an overall hypothesis, we need to take both observations into account.

3.3.1 Weighting

As mentioned above, there are various ways of integrating the mono-method hypotheses based on (a) reliability and (b) the effect size(s) of that predictor. We constructed possible overall hypotheses based on three ways of weighing reliability, expressed in the number of sources in which the predictor ($k = 1, \ldots, K$) appears ($a_k = 1,\ 2,\ 3$) and mean effect size ($b_k = 1, \ldots, 5$, representing a small to large effect, respectively). The mean effect sizes and the results of the different weightings, where the highest weight is the largest effect, can be found in Table 1. For W_1 the emphasis is on the reliability and for W_2 on mean effect size. In the first overall hypothesis (H_1) the ordering is established according to the mean effect size only, leading to.
Overall hypothesis 1:

$$\begin{aligned}&\{\beta_{\text{Feelings}},\ \beta_{\text{Type of trauma}},\ \beta_{\text{Safety}},\ \beta_{\text{Ordering chaos}},\ \beta_{\text{Trauma impact}},\ \beta_{\text{PTS (1, 3) months}}\}\\ &\quad>\{\beta_{\text{Parenting}},\ \beta_{\text{Coping}}\}\\ &\quad>\{\beta_{\text{Child characteristics}},\ \beta_{\text{Support}},\ \beta_{\text{Acute reactions}},\ \beta_{\text{Current outlook}},\ \beta_{\text{Heart rate}},\ \beta_{\text{Days in hospital}}\}\\ &\quad>\{\beta_{\text{Interpersonal relations}},\ \beta_{\text{Injury severity}}\} > \{\beta_{\text{Culture}},\ \beta_{\text{Trust}},\ \beta_{\text{Normalcy}},\ \beta_{\text{Gender}}\}.\end{aligned}$$

In the second overall hypothesis, the weight per determinant (W_{1k}) is established by:

$$W_{1k} = \max(b) \cdot a_k + b_k - \max(b),$$

leading to (H_2):

Overall hypothesis 2:

$$\begin{aligned}&\beta_{\text{Feelings}} > \beta_{\text{Parenting}} > \beta_{\text{Coping}}\\ &> \left\{\beta_{\text{Acute reactions}}, \beta_{\text{Child characteristics}}, \beta_{\text{Support}}\right\} > \left\{\beta_{\text{Interpersonal relations}}, \beta_{\text{Injury severity}}\right\}\\ &> \beta_{\text{Culture}} > \left\{\beta_{\text{Type of trauma}}, \beta_{\text{Safety}}, \beta_{\text{Ordering chaos}}, \beta_{\text{Trauma impact}}, \beta_{\text{PTS (1, 3) months}}\right\}\\ &> \left\{\beta_{\text{Current outlook}}, \beta_{\text{Heart rate}}, \beta_{\text{Days in hospital}}\right\} > \left\{\beta_{\text{Trust}}, \beta_{\text{Normalcy}}, \beta_{\text{Gender}}\right\}.\end{aligned}$$

And for the third hypothesis the weight per determinant (W_{2k}) is calculated by:

$$W_{2k} = \max(a) \cdot b_k + a_k - \max(a),$$

resulting in (H_3):

Overall hypothesis 3:

$$\begin{aligned}&\beta_{\text{Feelings}} > \left\{\beta_{\text{PTS (1, 3) months}}, \beta_{\text{Trauma impact}}, \beta_{\text{Type of trauma}}, \beta_{\text{Safety}}\ \beta_{\text{Ordering chaos}}\right\}\\ &> \beta_{\text{Parenting}} > \beta_{\text{Coping}} > \left\{\beta_{\text{Child characteristics}}, \beta_{\text{Acute reactions}}, \beta_{\text{Support}}\right\}\\ &> \left\{\beta_{\text{Current outlook}}, \beta_{\text{Heart rate}}, \beta_{\text{Days in hospital}}\right\} > \left\{\beta_{\text{Interpersonal relations}}, \beta_{\text{Injury severity}}\right\}\\ &> \beta_{\text{Culture}} > \left\{\beta_{\text{Trust}}, \beta_{\text{Normalcy}}, \beta_{\text{Gender}}\right\}.\end{aligned}$$

3.4 Step 4: evaluating the hypotheses

At this point we have established three mono-method hypotheses and three overall hypotheses based on (the integration) of different sources of empirical evidence. The empirical component of the children and trauma investigation ends here as new quantitative data need to be collected on all of the determinants in the overall hypotheses. However, as the aim of this paper is to introduce a method for integrating evidence from different sources, the final step of the proposed synthesis method can be demonstrated using hypothetical data. Note that the following statements cannot be used for inferences concerning the substantive field of children and trauma.

As we have a total of six possible hypotheses, the question is which hypothesis is most likely to be correct. For the worked example a three-stage procedure is suggested. In the first stage new quantitative data on all of the variables of the overall hypotheses is collected. This includes giving operational definitions of the qualitative predictors. In the second stage a statistical evaluation is made as to which overall hypothesis is the most likely. In the third stage a statistical evaluation is made as to whether the best overall hypothesis is more likely than the three mono-method hypotheses. Using this procedure will answer the questions of which determinants are the most important and whether an integration of evidence from different sources ultimately results in a better representation of the observed phenomenon.

The hypotheses stated above can be evaluated using a Bayesian model selection procedure especially developed for this purpose. This approach also allows for the testing of competing and contradicting hypotheses (Hoijtink et al. 2008). Note that other statistical procedures are also available. Results from this Bayesian analysis are expressed in a 'Bayes factor' (BF), which is a Bayesian model selection criterion. A BF is calculated for comparing two hypotheses and can be interpreted as the amount of support for one hypothesis over another hypothesis. A BF can be transformed into a posterior model probability (PMP) per hypothesis, and can be interpreted as an indication of which hypothesis is the most probable within the proposed set on a scale of 0–1.

Table 2 Results of hypothetical Bayesian evaluation of the hypotheses, with Bayes factors of hypothesis *i* against hypothesis *j* and the posterior model probabilities per hypothesis *i*

	$BF_{i,j}$					PMP_i
Step 1	i/j	H_1	H_2	H_3		
	H_1	1	0.2	4		0.23
	H_2	5	1	10		**0.71**
	H_3	0.25	0.1	1		0.06
Step 2	i/j	H_2	H_{MA}	H_{QS}	H_{EO}	
	H_2	1	10	5	6	**0.66**
	H_{MA}	0.1	1	0.5	0.4	0.06
	H_{QS}	0.2	2	1	2	0.16
	H_{EO}	0.25	2.5	0.5	1	0.12

In demonstrating our evaluation of the proposed method with hypothetical data, we assume that new (quantitative) data was collected on all 20 determinants. Following this, in stage 2, the best overall hypothesis needs to be found by evaluating them using the data collected in stage 1. The hypothetical results are presented in the upper panel of Table 2. As is shown by the boldfaced figure, the PMP is highest for the second overall hypothesis, PMP = 0.71. This can simply be interpreted as: Overall hypothesis 2 is the most likely hypothesis from all three overall hypotheses. Consequently, in stage 3, we compare Overall hypothesis 2 with the three mono-method hypotheses, resulting in the lower panel of Table 2. From the table we can conclude that the overall model is the most probable within this set of four hypotheses, PMP = 0.66. In this way we are able to say that integrating evidence from different sources results in a better representation of the observed phenomenon. This means that, in this case, the integration of different sources of evidence has resulted in a better representation of the phenomenon of interest.

4 Discussion

In this paper we have presented a method consisting of four steps for integrating evidence from different sources. The core of this method concerns the integration of findings of a different nature by translating them into (directional) hypotheses. The resulting overall directional hypotheses can be considered formulaic representations of a phenomenon. In other words, the method provides a condensed theory of a phenomenon and can be statistically tested when confronted with new empirical data.

We see three major advantages in the method that is presented here. First, the method is capable of handling different sources of evidence. As in previous findings (e.g., Dixon-Woods et al. 2006a), our worked example demonstrated that different sources are complementary; the determinants found in the qualitative synthesis, the meta-analysis and the expert elicitation did not completely overlap and effect sizes were different. This makes integration of different sources a worthwhile effort.

Second, the objective of achieving the overall hypothesis is not limited to, for example, the formulation of a program theory which underlies an intervention, such as in a realist synthesis (Pawson et al. 2004). The method we propose can also generate epidemiological models involving determinants influencing a phenomenon, as was shown in the substantive case.

Third, we consider the possibility of evaluating and testing the outcomes of the whole integrative endeavor to be a strong feature. Many synthesis efforts can be tracked down by

their narrative accounts of the systematic procedure that was followed, but few can test the outcomes with empirical data as we propose in the final step. Consequently, we are able to check how well the overall hypothesis, i.e., the newly constructed model, fits social reality. Furthermore, due to the Bayesian model selection approach we can, when necessary, compare several contradicting hypotheses (Hoijtink et al. 2008).

Some elements of the proposed method need further development. First, the review question that guided the example in this paper concerned finding determinants and therefore fits the relatively simple multiple regression model. In the case of more complex review questions, e.g., involving more dependent variables and indirect effects, the directional hypotheses would also be more complex. Consequently, the statistical methods for evaluating them (e.g., directional structural equation models) need to be further developed. Fortunately, statistical techniques to evaluate (multivariate) analysis of variance models and repeated measurements already exist (Mulder et al. 2009).

A second challenge is the comparison of variables (quantitative) and themes or concepts (qualitatively). It is in the nature of quantitative research that variables be operationally defined so that they can be measured. Validated measuring instruments can often be used. In contrast, qualitative methodologies use concepts that have no fixed content, as it is the aim of the research itself to discover relevant themes and define concepts in a way that fits the field of research. When using these concepts in statistical analyses of empirical data, they need to receive operational definitions. Also note that the proposed method is appropriate for themes and concepts in qualitative studies that are quantifiable to a certain extent. Our further research will focus on transforming qualitative findings of a different nature, such as views and experiences. This is already difficult in itself, but adding to its complexity is the fact that concepts might be used in dissimilar ways in studies using different methodologies. Closely related to this issue is the fact that qualitative and quantitative studies on a certain topic differ too much with regard to the findings and no overlap exists. Although this is a finding in itself it leaves little to be synthesized.

A third challenge has to do with weighing the findings of the different sources. This can be done in an abundance of ways and on multiple levels; (1) weighing qualitative concepts in order to determine their effect sizes, might result in bias. This might be due to research trends (some topics being more popular than others), to over- or under-representation of certain research disciplines, and obvious or in-depth research findings, respectively, and (2) weighing the mono-method hypotheses when integrating them, where the particular weights of the qualitative and quantitative findings in relation to each other need to be further investigated, for instance by using sensitivity analyses.

Finally, as is the case for all theoretical models, the hypotheses grounded in the different sources are only a simplified version of reality. Consequently, we pay for simplicity with information loss. More specifically, by using Cohen's rules of thump (small, medium, or large effects) for interpreting the weighted mean effect sizes of the quantitative determinants (found in the meta-analysis) we lose precision we originally possessed. Along the same line, establishing comparable 'effect sizes' for qualitative findings across clinical and contextual settings needs some refinement in future research.

The method described here can be roughly divided in two components: (1) evidence synthesis and (2) model evaluation. The actual synthesis of the evidence occurs in evidence synthesis (1). In model evaluation (2) the hypothesis is tested using new empirical data and can be considered a modest way of validating the generated theoretical model. It provides additional insights into the correctness of the generated hypothesis. Still, both components can be used as stand-alone approaches. The results of evidence synthesis (1) can provide a program theory, whereas the results of model evaluation (2) can tell us how well this

theory explains social reality. In summary, the complete approach can be used to develop and evaluate new interventions or other theoretical models.

We believe the beauty of our method to be that it truly integrates different types of evidence while treating each source as equally important. It facilitates the processing of an extensive amount of studies, while at the same time the outcome is *one* or *a few* hypotheses, which adds to the elegance of the method. The results generated by using this method can be used for the development or refinement of program theories or other theoretical models. The method needs further refinement and evaluation in different fields to establish its use and to reflect upon its contributions to the field of evidence-based policy and practice.

Acknowledgments This research was supported by Grant NWO-VICI-453-05-002 of the Netherlands Organization for Scientific Research.

Appendix

To investigate the views of experts in the field of children and trauma, a purposive sample six Dutch mental health care professionals were interviewed. Participants were emailed for their cooperation. Each interview (performed by the first author) was held at the expert's office and lasted about an hour. The interviews were audio-recorded and transcribed and field notes were made by the interviewer. The interview questions were open-ended and involved personal acquaintance, introduction about the research (including discussing confidentiality), and uncovering the determinants of PTSD development and their relative importance.

The transcribed interviews and the field notes were analyzed using computer software for qualitative analysis (QSR NVivo 8). The data were open-coded. Codes were discussed among the first two authors and agreed upon. Next, axial and selective coding were performed, resulting in the following themes: type of trauma (single or multiple trauma, human against human violence or natural disasters, amount of loss of control), severity (level of gravity of trauma experienced), ordering chaos (ability to reorganize chaos of feelings, experiences and memories), feelings (e.g., guilt, shame, loneliness, fear, anger), safety (sense of safety experienced), parenting (parents' handling of child in terms of protection, help coping, exemplifying healthy reaction and availability), avoidance (ability to avoid situations reminding of traumatic event), child characteristics (age, temperament, cognitive/social intelligence, development and self-image), support (help offered and understanding shown by friends, family and community), trust (amount of faith in surrounding people), care after trauma (consolation received during/shortly after trauma), culture (collectiveness of community and it's conventions of trauma), interpersonal relationships (interactions with friends, teachers and family).

References

Alisic, E., Jongmans, M.J., van Wesel, F., Kleber, R.J.: Risk and protecting factors for posttraumatic stress in children: a systematic review of prospective studies. Clin. Psychol. Rev. **31**, 736–747 (2011)

Barbour, R.S.: The role of qualitative research in broadening the 'evidence base' for clinical practice. J. Eval. Clin. Pract. **6**(2), 155–163 (2000)

Boeije, H.: Analysis in Qualitative Research. Sage, London (2009)

Candy, B., King, M., Jones, L., Oliver, S.: Using qualitative synthesis to explore heterogeneity of complex interventions. Med. Res. Methodol. **11**, 124 (2011)

Cohen, J.: A power primer. Psychol. Bull. **112**, 155–159 (1992)

Dixon-Woods, M., Fitzpatrick, R., Roberts, K.: Including qualitative research in systematic reviews: opportunities and problems. J. Eval. Clin. Pract. **7**(2), 125–133 (2001)
Dixon-Woods, M., Agarwal, S., Jones, D.R., Young, B., Sutton, A.J.: Synthesising qualitative and quantitative evidence: a review of methods. J. Health Serv. Res. Policy **10**, 45–53 (2005)
Dixon-Woods, M., Bonas, S., Booth, A., Jones, D.R., Sutton, T.M.A.J., Shaw, R.L., et al.: How can systematic reviews incorporate qualitative research? A critical perspective. Qual. Res. **6**(1), 24–44 (2006a)
Dixon-Woods, M., Cavers, D., Agarwal, S., Arthur, E.A.A., Harvey, J., Hsu, R., et al.: Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. Med. Res. Methodol. **6**, 35 (2006b)
Gorecki, C., Brown, J.M., Nelson, E.A., Briggs, M., Schoonhoven, L., Dealey, C., et al.: Impact of pressure ulcers on quality of life in older patients: a systematic review. JAGS **57**, 1175–1183 (2009)
Hannes, K., Macaitis, K.: A move to more transparent and systematic approaches of qualitative evidence synthesis: update of a review on published papers. Qual. Res. **12**(4), 402–442 (2012)
Harden, A., Thomas, J.: Methodological issues in combining diverse study types in systematic reviews. Int. J. Soc. Res. Methodol. **8**(3), 257–271 (2005)
Harden, A., Brunton, G., Fletcher, A., Oakley, A.: Teenage pregnancy and social disadvantage: systematic review integrating controlled trails and qualitative studies. BMJ **339**, 4254 (2009)
Hoijtink, H., Klugkist, I., Boelen, P.A. (eds.): Bayesian Evaluation of Informative Hypotheses. Springer, New York (2008)
Hunter, J.E., Schmidt, F.L., Jackson, G.B.: Meta-analysis: Cumulating Research Findings Across Studies. Sage, Beverly Hills (1982)
Lipsey, M.W., Wilson, D.B.: Practical Meta-analysis. Sage, Newbury Park (2001)
Mulder, J., Hoijtink, H., Klugkist, I.: Inequality and equality constrained multivariate linear models: objective model selection using constrained posterior priors. Stat. Plan. Inference **140**, 887–906 (2009)
Mulder, J., Hoijtink, H., de Leeuw, C.: BIEMS: a Fortran90 program for calculating Bayes factors for inequality and equality constrained models. J. Stat. Softw. **46**(2), 1–39 (2012)
Noblit, G.W., Hare, R.D.: Meta-ethnography: Synthesizing Qualitative Studies. Sage, London (1988)
Paterson, B.L., Thorne, S.E., Canam, C., Jillings, C.: Meta-study of Qualitative Health Research: A Practical Guide to Meta-analysis and Meta-synthesis. Sage, Thousand Oaks (2001)
Pawson, R., Greenhalgh, T., Harvet, G., Walshe, K.: Realist Synthesis: An Introduction. University of Manchester, Manchester (2004)
Pope, C., Mays, N., Popay, J.: Synthesizing Qualitative and Quantitative Health Research: A Guide to Methods. Open University Press, Maidenhead (2007)
Roberts, K.A., Dixon-Woods, M., Fitzpatrick, R., Abrams, K.R., Jones, D.R.: Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. Lancet **360**, 1596–1599 (2002)
Sandelowski, M., Barroso, J.: Handbook for Synthesizing Qualitative Research. Springer, New York (2007)
Sandelowski, M., Voils, C.I., Barroso, J.: Defining and designing mixed research synthesis studies. Res. Sch. **13**(1), 29–40 (2006)
Sandelowski, M., Voils, C.I., Barroso, J.: Comparability work and the management of difference in research synthesis studies. Soc. Sci. Med. **64**(1), 236–247 (2007)
Sherwood, G.: Meta-synthesis: merging qualitative studies to develop nursing knowledge. Int. J. Hum. Caring **3**, 37–42 (1999)
Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., et al.: Integrating qualitative research with trials in systematic reviews. BMJ **328**, 1010–1012 (2004)
van Wesel, F., Hoijtink, H., Klugkist, I.: Choosing priors for inequality constrained normal linear models: methods based on training samples. Scand. J. Stat. **38**, 666–690 (2010)
van Wesel, F., Boeije, H., Alisic, E., Drost, S.: I'll be working my way back: a qualitative synthesis on the trauma experiences of children. Psychol. Trauma Theory Res. Pract. Policy (2011). doi:10.1037/a0025766