This paper is published as

Verhage, A. & Boels, D. (2016). “Critical appraisal of mixed methods research studies in a systematic scoping review on plural policing”, *Quality & Quantity,* Online first, DOI 10.1007/s11135-016-0345-y.

The final publication is available at:

[http://link.springer.com/article/10.1007/s11135-016-0345-y](http://www.springer.com/-/0/AVSNUHQP2brxj7RS5XrD)

**Critical appraisal of mixed methods research studies in a systematic scoping review on plural policing.**

Assessing the impact of excluding inadequately reported studies by means of a sensitivity analysis.

**Key terms** scoping review, systematic review, critical appraisal, sensitivity analysis, meta-integration, thematic synthesis, mixed methods synthesis

**Abstract**

A scoping review is a method which is often applied in the systematic review arena. Its aim is to map existing literature in a specific area. Carrying out a scoping review entails dealing with a number of methodological questions that arise during the review. In this paper, we go into the way in which these issues were contemplated upon and dealt with. We particularly focus on how to deal with the critical appraisal of literature (and the use of the outcome of this appraisal in terms of in- or exclusion criteria) and its effects on the contents and outcome of the review. This implies that we will report on the result of a sensitivity analysis. A second methodological issue that is focused on is how to make a review that considers multiple (qualitative, quantitative and mixed methods) designs and how to handle the different types of results that were derived from each method. We conclude by considering the pros and cons of including inadequately reported studies and plea to inform readers on the quality of literature that was included.

1. **Introduction**

Literature reviews can be carried out in different manners and by making use of several systematic approaches (Booth, 2012). Although it might seem that systematic reviews are relatively clear-cut in terms of criteria-building, selection and reporting, in literature a number of issues remain unclear. In a review study with regard to *plural policing[[1]](#footnote-1)*, the authors came across a number of these problematic choices and decided to report on the way that they have dealt with these choices and what the effects of these choices are. Two of the main problems were (1) how to deal with studies that report very minimally on the used methodology (critical appraisal) and the effects of in- and exclusion and (2) how to work with quantitative, qualitative and mixed methods research evidence.

In this paper we will show that the combination of these two issues has steered us to developing an original (and pragmatic) methodological approach that involves both the inclusion of multiple methodologies (quantitative, qualitative and mixed methods) and the decision on how to deal with so-called lower-quality studies.

The study involved concerns a **scoping review**, a method to map the existing literature in a broad thematic area (Pham et al., 2014). More specifically, a scoping review aims to “*map the literature on a particular topic or research area and provide an opportunity to identify key concepts, gaps in the research, and types and sources of evidence to inform practice, policymaking and research*” (Daudt, van Mossel & Scott, 2013, 8). It can be understood as an assessment of the size and scope of available research literature (Booth, 2012, 27). Scoping reviews typically entail at least five key stages: (1) identifying the research question, (2) identifying relevant literature, (3) study selection, (4) charting the data, (5) collating, summarizing and reporting the results (Arksey & O’Malley, 2005). With regard to the third phase, study selection, there is an ongoing debate regarding the **need for quality assessment** or critical appraisal of primary studies (Pham et al., 2014). A critical appraisal “*seeks to assess the validity and reliability of a primary research study and its findings*” (Carroll & Booth, 2015, 149). Whereas the ‘founding fathers’ of the scoping review methodology state that “*quality assessment does not make part of the scoping (review) study remit*” (Arksey & O’Malley, 2005, 22), other authors stress the importance of a thorough quality assessment in scoping reviews (Daudt et al., 2013). This debate has also been held in the literature on systematic reviews, particularly regarding qualitative research evidence. Although the literature points to an increasing consensus to conduct a quality assessment of qualitative studies in evidence synthesis, some debate still exists about how such an appraisal should best be done (Carroll & Booth, 2015). Recent literature suggests that reviewers need to choose the quality criteria in function of the aims and questions of the review (Hannes, 2011; Toye et al., 2013). In this respect, the adequateness of quality criteria will depend upon the context of the review. Irrespective of what criteria are used to assess the quality of the primary studies, the outcome will be that some studies are considered to be of ‘high quality’, whereas others will be labelled as ‘low(er) quality’ studies. The question then poses itself: what should the outcome of this label be? Should these studies be systematically excluded from the review? Should we give them lower weight (but how?)? This question has been dealt with earlier in other studies, but the actual potential effects of each option for the content and outcome of the review in question is difficult to estimate. One possible way of dealing with this is to conduct a post hoc sensitivity analysis to explore the impact of excluding studies below a certain quality threshold on the original synthesis results (Boeije, van Wesel & Alisic, 2011). In this article, we aim to contribute to this debate by reporting on the results of a sensitivity analysis in our scoping review on plural policing.

A second point of discussion in literature on (systematic) reviews concerns the synthesis of qualitative, quantitative and mixed-methods research evidence (Frantzen & Fetters, 2015). Synthesis of *quantitative* evidence usually occurs by means of a meta-analysis (Green et al., 2008; Paterson, 2012). Currently, several methods exist to synthesize *qualitative* research evidence[[2]](#footnote-2) (for an overview, see [Dixon-Woods, Agarwal, Jones, Young, & Sutton, 2005](#_ENREF_6); [Hannes & Lockwood, 2012](#_ENREF_16); [Manning, 2012](#_ENREF_21); [Noyes, Popay, Pearson, Hannes, & Booth, 2008](#_ENREF_23); [Paterson, 2012](#_ENREF_24); [Ring, Ritchie, Mandava, & Jepson, 2010](#_ENREF_28)). In recent years, much advancement has been made in this field. However, less practical aid is found in literature with regard to synthesis of qualitative, quantitative and mixed methods research findings. To our knowledge, only one study offers detailed guidance to this end (Frantzen & Fetters, 2015). In this paper, we offer support for and discuss the use of such a synthesis method, namely basic convergent meta-integration. In this way, we hope to contribute to the body of literature on integrating qualitative, quantitative and mixed methods findings, or what Frantzen & Fetters (2015, 2) refer to as “*mixed studies reviews*”.

By focusing on these two issues that we were confronted with in our scoping review, this paper has a threefold aim. First, we aim to contribute to the literature on critical appraisal in reviews by carrying out a sensitivity analysis starting from a different point of view: a scoping review on plural policing, including qualitative, quantitative and mixed-methods research findings. As such, the article goes beyond original methodological approaches, as it is based on a scoping review, including findings from qualitative, quantitative and mixed-methods primary research. Secondly, this paper aims to contribute to the debate on the synthesis of qualitative, quantitative and mixed methods primary findings, by providing insights into a recently conducted synthesis based on Frantzen & Fetter’s (2015) basic convergent meta-integration. Lastly, this paper will add to the existing methodological framework on scoping reviews by illustrating the importance of quality assessment and sensitivity analyses in scoping reviews (Pham et al., 2014).

1. **How to deal with critical appraisals when reviewing**

A critical appraisal implies that studies are screened with regard to their quality. According to Dixon-Woods, Booth and Sutton (2007), the findings from a critical appraisal should be used in a meaningful way. Hannes (2011) differentiates between three possible alternative outcomes for lower quality studies. A first possibility is to exclude lower quality studies from the evidence synthesis. Some reviewers explain that excluding qualitative research based solely on design holds the risk to miss insights relevant for a good understanding of the phenomenon ([e.g. Booth, 2001](#_ENREF_1)). As stated, a possible way to deal with this is by means of a post hoc sensitivity analysis (Boeije, et al., 2011). This is a key point of discussion in the current methodological literature (Carroll, Booth & Lloyd-Jones, 2012). Previous studies have already carried out sensitivity analyses (e.g. Carroll, et al., 2012; Franzel, Schwiegershausen, Heusser & Berger, 2013; Noyes & Popay, 2007; Thomas & Harden, 2008) and most have found that excluding inadequately reported studies or lower quality studies did not significantly affect the results of the synthesis. Franzen and colleagues, conducting a meta-ethnography, assessed reporting criteria, criteria related to the plausibility and coherence of the results and reflexivity. They conclude that “*the themes from the excluded papers would not have altered the meta-synthesis*” (Franzen et al., 2013, 7). Noyes and Popay (2007) and Thomas and Harden (2008) equally concluded on the basis of their sensitivity analysis that lower quality studies had no significant impact on qualitative reviews. Noyes and Popay (2007, 230), looking at the thickness of the descriptions and the technical quality of the application of the methods, concluded that “*we undertook an analysis of whether anything substantially different was found in weaker studies, which it was not*”. Thomas and Harden (2008, 45) concluded on the basis of their sensitivity analysis - based on reporting criteria, on the sufficiency of the strategies and the appropriateness of the study methods - that “*the poorer quality studies contributed comparatively little to the synthesis and did not contain many unique themes*”.

All in all, previous research suggests that the exclusion of ‘lower quality studies’ does not significantly impact on the synthesis results and that such studies tend to contribute less to the synthesis than higher quality studies (Carroll & Booth, 2015). The problem with this conclusion, however, is that previous research has used different critical appraisal criteria and different synthesis methods. Therefore, more research is needed to test the value of sensitivity analysis and to determine its applicability to different types of syntheses, evidence or questions (Carroll & Booth, 2015, 153). To add to this body of research, in this paper we therefore report on the results of our sensitivity analysis.

A second possibility is to weigh the evidence, to give more weight to studies that score high on quality. Boeije et al. (2011) explored a method of weighting studies and their contribution to a synthesis. Using scores achieved by checklists and expert judgement, they offered more weight to findings from higher quality studies. They found that weighting studies did change their synthesis results, in that the order of themes organized by amount of evidence would change because of the weighting with the quality appraisal. These first two options are based on the rationale that studies of insufficient quality may distort the synthesis and may cause difficulties in interpretation (e.g. [Dixon-Woods, Fitzpatrick & Roberts, 2001](#_ENREF_9); [Dixon-Woods, Shaw, Agarwal & Smith, 2004](#_ENREF_10)).

A final option is to simply describe what has been observed without excluding any studies. This last option is not recommended by the Cochrane Qualitative Research Methods Group (Hannes, 2011).

1. **Scoping reviews and critical appraisal**

As mentioned in the introduction, the main aim of a scoping review is to identify the extent and nature of research evidence (Booth, 2012, 27). It is a systematic approach to a literature review and can be divided into several steps. In previous work (Verhage and Boels 2015), we have summarized the different steps in a *systematic* review based on a literature review (see Figure 1). In comparison to the key phases of a scoping review (see introduction), we feel the steps for a systematic research offer slightly more detailed grip to conduct a literature search. Therefore, we followed the steps of a systematic review in our scoping on plural policing.

In the following sections of this paper, we discuss these steps, the challenges we encountered during our own review and some strategies to confront the challenges.

Figure 1: Steps in a systematic review



**3.1 Research protocol**

A review includes a research plan ([Pearson, 2004](#_ENREF_25)), a detailed description of the research questions or hypotheses ([Hannes & Claes, 2007](#_ENREF_15)), and eligibility criteria or the criteria to include studies in the review ([Green et al., 2008](#_ENREF_13)). Our research protocol included three initial research questions:

* + RQ 1: What are the dangers of blurring boundaries and in which contexts or cases are they recognized?
	+ RQ2: What are the effects of plural policing on ‘core tasks’ of the public police with regard to the equal division of safety and security in society?
	+ RQ3: What are the differences between policing actors regarding the use of discretionary space and how does this affect citizen’s legal recourses?

Simultaneously, five inclusion or eligibility criteria were formulated: empirical research[[3]](#footnote-3), link with research questions, research published/conducted between 1990-current (plural policing as a concept only acquired a central position in the criminological literature since the nineties ([Terpstra & Stokkom, 2015](#_ENREF_32))), published and unpublished studies (in order to cover as much relevant literature as possible ([Hammerstrom, Wade, & Klint Jorgensen, 2010](#_ENREF_14); [Hannes & Claes, 2007](#_ENREF_15)), studies written in English and Dutch. The last criterion relates to our aim to disclose research in Dutch to a non-Dutch speaking research network. As a result, studies with different contexts were included in the review (Hannes & Harden, 2011). We also formulated exclusion criteria. For instance, we excluded studies focused on partnerships in which the steering function and main policing functions remain in hands of the public police. We did not exclude studies with specific research designs, but searched across the entire range of empirical research (qualitative and quantitative).

During our review, we encountered some issues with the predetermined research questions (Dixon-Woods et al., 2006). More specifically, hardly any studies addressed the third research question, which is why the research team (consisting of three researchers) decided to drop that question. Furthermore, studies hardly addressed our research questions directly, which is why we included studies of which we thought they addressed the two research questions indirectly. This implies a subjective assessment in the early stage of the review. Regular discussion between the members of the research team could confirm inclusion of studies, but could not assure that *all* publications which also addressed the research questions indirectly were included. We did not choose to reformulate our first two research questions, but we divided the first research question in two sub-questions during our search process. This was based on the findings that several studies provided some information on the potential downsides of plural policing in se, more than of blurring boundaries between policing agents. Altering the research question in the middle of the review entailed the risk of having missed important studies and having included irrelevant studies, which in turn would have required to start the whole selection process over.

**3.2 Search strategy**

In systematic searches for literature, two main approaches exist towards the search strategy: a comprehensive search or a selective search. A comprehensive search typically identifies and includes all relevant studies ([Hammerstrom et al., 2010](#_ENREF_14); [Hannes & Claes, 2007](#_ENREF_15); [Pearson, 2004](#_ENREF_25)). Whereas a comprehensive search strategy has important advantages ([e.g. Manning, 2012](#_ENREF_21)), some authors argue that saturation prevails on quantity, thereby favouring a selective search strategy using sampling techniques (e.g. [Booth, 2001](#_ENREF_1); [Noyes et al., 2008](#_ENREF_23)). In the end, which search strategy would finally be chosen (comprehensive of selective), will depend on the philosophical/epistemological vision of the researcher and the goals of the review ([Ring et al., 2010](#_ENREF_28)). Given our critical realist epistemological position and our focus on summarizing data more than developing new concepts or theories, we aimed for a comprehensive linear search strategy (Barnett-Page & Thomas, 2009).

Based on a general literature review on plural policing, we determined key words and databases to identify relevant publications. Studies in English were searched on the basis of 12 keywords[[4]](#footnote-4), used in 16 databases (six general databases for published research, three for grey literature and seven policing journals). Initially, it was our intention to additionally search unpublished research by browsing the program booklets of the annual conferences of the European Society of Criminology (ESC) and the American Society of Criminology (ASC). Time constraints made us refrain from doing so. Studies in Dutch were identified on the basis of ten keywords[[5]](#footnote-5), used in 13 databases (six general databases for published research, four relevant journals, one publisher’s website, two databases of research institutes).

**3.3 Selection: from longlist to shortlist**

We meticulously registered the total amount of publications found per keyword per database, the number of publications that did not address the topic of plural policing and the number of publications that were eligible for the longlist[[6]](#footnote-6). In other words, this first selection process – based on titles (Papaioannou, 2012) (or keywords/abstracts if titles gave no hits) - excluded all studies that did not address the topic of plural policing or clearly did not address any of our three research questions. This strategy resulted in a longlist of 707 studies for the ‘English search’. After eliminating all double references (which led to a final longlist of 308 studies), a second selection process took place (Pearson, 2004). Full reading of abstracts and/or quick scan of all unique publications on the longlist enabled us to eliminate publications that did not meet our inclusion criteria (Papaioannou, 2012). Of the 308 studies in the longlist, 47 studies were withheld as potentially entirely meeting our inclusion criteria. Based on full reading of these 47 articles/chapters/books, the final shortlist was established, which contained 25 publications that fully met our inclusion criteria. A schematic overview is found in Figure 2.

The longlist of publications found through the ‘Dutch search’ contained 160 publications, which was reduced to 138 after eliminating double references. The further selection process followed the same steps as the one for the English publications and left us with six publications that fully met our inclusion criteria. In total, we thus retained 31 publications for further analysis. Of these 31 publications, 23 were based on a qualitative design, four on a quantitative design and four[[7]](#footnote-7) made use of a mixed design, including both qualitative and quantitative methods of data collection.



**3.4 Analysis of the primary studies: quality assessment and synthesis**

*3.4.1 Quality assessment*

In general, there is an increasing consensus about the importance of assessing the quality of primary studies, at least in systematic reviews (Toye et al., 2013). However, there is still debate about the methods used to assess quality and the implication of the appraisal (Carroll & Booth, 2015; [Dixon-Woods et al., 2006](#_ENREF_7)).

The latest literature on *systematic* reviews indicates the importance of conducting a critical appraisal of the quality of primary studies (Carroll & Booth, 2015; Toye et al., 2013). Hannes and Macaitis (2012) analysed 82 qualitative evidence syntheses published between 2005 and 2008 and found that the majority of the reviews included a critical appraisal. Different guidelines and checklists exist to assess *qualitative* studies for their quality[[8]](#footnote-8) ([Dixon-Woods et al., 2001](#_ENREF_9); [Dixon-Woods et al., 2004](#_ENREF_10); [Hannes, Lockwood & Pearson, 2010](#_ENREF_17); [Manning, 2012](#_ENREF_21); [Noyes et al., 2008](#_ENREF_23); [Spencer, Ritchie, Lewis & Dillon, 2003](#_ENREF_30)). Even though such checklists can be valuable, there are still some notes and critical remarks for these guidelines and checklists. Firstly a consensus concerning what criteria are adequate and should thus be included in the checklists does not yet exist ([Dixon-Woods et al., 2001](#_ENREF_9)). It also remains unclear what checklist should preferably be used ([Manning, 2012](#_ENREF_21)). Furthermore, a comparative study of three checklists that allow a structured assessment from individual studies indicated that the three lists did not lead to a greater consistency in decisions concerning the selection of studies ([Dixon-Woods, 2007 in Noyes et al., 2008](#_ENREF_23)). [Ring et al. (2010)](#_ENREF_28) add the concern that the use of checklists could reduce qualitative research to merely executing technical procedures. Also, supporters of the checklists warn for a too rigid use of them (e.g. [Hannes et al., 2010](#_ENREF_17); [Spencer et al., 2003](#_ENREF_30)). [Dixon-Woods et al. (2004)](#_ENREF_10) also state that a lot of checklists are based on the incorrect idea that qualitative research is uniform in terms of methods for data collection an analysis. They argue that qualitative research only has a few common (methodological) characteristics and propose a number of ‘prompts’ to evaluate qualitative research on these points. As these factors are general, they can be completed with evaluation criteria specific for certain methods ([Dixon-Woods et al., 2001](#_ENREF_9)). Other researchers are also of the opinion that it could be interesting to premise general criteria to evaluate qualitative studies ([see e.g. Popay, Rogers & Williams, 1998](#_ENREF_26)).

In most *scoping* reviews, no quality assessment of primary studies is performed (Booth, 2012, 19; Pham et al., 2014, 376). However, following our critical realist stance (Barnett-Page & Thomas, 2009), we stress the importance of a quality assessment of primary studies. In our original review, we intended to make use of the generic prompts proposed by [Dixon-Woods et al. (2004)](#_ENREF_10) to assess the overall quality of the studies featuring on our shortlist (see Figure 4). The choice for this instrument was in part steered by the fact that we preferred one instrument that could be applied to all qualitative, quantitative and mixed methods research. Therefore, we chose an instrument that is sufficiently general and not applicable to only one type of research.

Figure 4: Prompts for assessing quality in primary research (Dixon-Woods et al., 2004, p. 224)

However, the lack of methodological information in many primary studies hampered the appraisal of these general prompts. Clearly, the biggest obstacle we faced in the quality appraisal of the studies was the limited information on the methodology. Although most - but not all - studies directly[[9]](#footnote-9) reported the methods of data-collection, a great deal of variation was found regarding the extent to which they report other methodological aspects (e.g., research questions, sampling, data-analysis). For instance, 19 (of the 31) studies did not mention research questions. Information on sampling (why and how respondents/units are sampled) was missing in 14 studies, but some information regarding the sample - for instance a description of the sample - was found in seven publications. Due to lack of information, it was not possible to assess the quality of data analysis in 25 studies. Therefore, the appraisal of several studies included in our review unfortunately turned out somewhat disappointing: most studies simply could not be assessed on the general prompts. As a result, we did not exclude these studies from our review. This choice was based on our aim to conduct a scoping review and to have an overview of what can be found in the empirical literature regarding the dangers of plural policing and the impact on police core tasks. Given the limited direct information, we certainly did not want to exclude the little relevant information available. After all, if we decided to follow the appraisal result, we would have had to exclude 25 out of 31 studies, resulting in a very limited view on the scope of research in this field. We were also not convinced of the conclusion that these studies were of low quality – we could after all not assess whether there were methodological flaws in the publication of in the entire study in itself. Although this is a valid choice, we do wonder if we had obtained different end results had we excluded ‘low quality’ studies. Therefore, we conducted a sensitivity analysis, of which the results are reported below. Before we go into the results of this analysis, we give a brief overview of how the findings from primary studies were synthesized.

*3.4.2 Synthesis of research evidence: basic convergent meta-integration*

Quantitative research evidence is usually synthesized by means of a meta-analysis ([Green et al., 2008](#_ENREF_13); [Paterson, 2012](#_ENREF_24)). To summarize qualitative evidence, several methods exist ([Dixon-Woods et al., 2005](#_ENREF_6); [Paterson, 2012](#_ENREF_24); [Ring et al., 2010](#_ENREF_28)). However, less practical aids exist regarding the synthesis of evidence springing from qualitative, quantitative and mixed-methods research. Recently, Frantzen and Fetters (2015) presented different types of meta-integration that allow for such synthesis. They define meta-integration as “*the combining and contrasting of results from quantitative, qualitative and mixed-methods studies (as applicable) to identify patterns, relationships, and disagreements, and/or to develop deeper insights and interpretations by bringing together and integrating all relevant research findings, regardless of the type of data, for a deeper understanding of a phenomenon*” (Frantzen & Fetters, 2015, 22). In our review, we followed a procedure closely linked to what Frantzen and Fetters (2015, 12) call basic convergent meta-integration, in which quantitative findings are transformed into qualitative findings. In essence, this means that quantitative variables were converted into qualitative themes by utilizing descriptive conclusions from the quantitative papers without using the numeric results (Frantzen & Fetters, 2015). Thus, we did not use the numeric results from the quantitative (and mixed methods papers), but only coded text. As the findings of the mixed methods papers were presented separately (i.e. the quantitative and qualitative results were reported separately), we could easily transform the quantitative findings into qualitative findings (Frantzen & Fetters, 2015). We did not synthesize the findings from the qualitative and transformed quantitative and mixed methods papers separately (intra-method analysis and synthesis, Frantzen & Fetters, 2015, 15). Instead, we synthesized all papers simultaneously. This was done by conducting thematic synthesis as developed by Thomas and Harden (2008).

Regarding the free line-by-line coding, we adopted an integrative approach, meaning that we looked at the data with our research questions in mind and only coded text fragments that were relevant for the research questions ([Cruzes & Dyba, 2011](#_ENREF_4)). We did not start the coding process with a specific coding framework in mind. As such, we did not code line-by-line (but only coded relevant fragments), which is not so uncommon when large number of studies are included ([Cruzes & Dyba, 2011](#_ENREF_4)). As mentioned above, the same approach was used to code findings obtained from quantitative studies and mixed methods studies (Frantzen & Fetters, 2015). In these studies, the non-numerical data (text) were coded. The only difficulty we experienced in this phase is one that has been described in previous research, namely the difficulty to determine what counts as data or findings in qualitative papers ([Thomas & Harden, 2008](#_ENREF_33)). We coded text fragments in all sections of the studies, but only included them for further analysis if they were clearly inferred from the empirical study. We found that these mainly stemmed from the results and discussion sections. Inferences from authors with less clear links to the empirical results (mostly in the discussion section) were included - if relevant for the research questions - but summarized under a separate theme.

The initial codes were subsequently processed to eleven descriptive themes (e.g. conflict of interest, misconduct). This thematic synthesis was first conducted for the ‘English’ studies. After having constructed the descriptive themes, we line-by-line coded the ‘Dutch’ studies and analysed the extent to which the free codes could be integrated in the descriptive codes derived from the English studies. We found that we did not have to construct new descriptive codes as all free codes could be integrated in the ‘English’ descriptive codes.

We deliberately did not proceed to the third stage, namely the development of analytical themes. Although several authors argue that this comes down to ‘re-codifying the known’ ([Finfgeld-Connett, 2014, p. 6](#_ENREF_12)) and urge to move beyond mere reorganisation and re-categorisation of existing qualitative research findings, several reasons underlay our choice. Firstly, the third step entails the construction of new concepts, hypotheses or theories, which suggests that it is of greater relevance for an interpretative synthesis than for an integrative synthesis. Our aim was not to develop new understandings about the dangers of plural policing or about the impact of plural policing on the core tasks of the public police, but to offer a thorough overview of the empirical literature on these matters (cf. scoping review). Secondly, the apparent lack of direct focus on our research questions in primary studies and thus the limited available relevant empirical data also refrained us from going beyond the empirical findings ([Finfgeld-Connett, 2014](#_ENREF_12)). Although [Thomas and Harden (2008)](#_ENREF_33) suggest that conceptual innovation is necessary just because primary studies do not address research questions directly, we feel that developing analytical themes would have been too risky as our inferences and conclusions would not be sufficiently empirical-based. Our choice not to go beyond the primary data is also inspired by the high variation in research contexts of the primary studies: they were characterised by highly different research settings, topics and aims[[10]](#footnote-10). The nature of plural policing also influenced our limitation to descriptive themes: activities and authorities of the public police and non-public policing actors can be so different in different (national) settings that it is necessary to retain the context. For instance, the role and authorities of public police in fragile African states is so different than in the UK that making abstractions - and thus loosing context - becomes meaningless. A potential way to overcome such issues is to conduct a context-specific instead of a multi-context review (Hannes & Harden, 2011). However, given the limited research focused on our two research questions, focusing on only one context would not have resulted in a sufficient number of studies to conduct a synthesis providing meaningful results or insights.

*3.4.3 Sensitivity analysis*

In our original scoping review, we included all 31 studies, irrespective of their quality. However, in this section we explore the extent to which exclusion of ‘lower quality’ studies impacts on the original synthesis results. Following Carroll et al., (2012), our sensitivity analysis is based on reporting criteria: we assess whether excluding studies that inadequately report their methods demonstrate any effect on the findings of our thematic synthesis. Thus, we did not assess the actual conduct of the study or the appropriateness of the methodology, but only focused on *how the methodology is described in the primary studies*. The choice for using reporting criteria is based on two reasons. First, having methodological information is vital in order to assess the quality: one can only assess what is reported in a publication. This is what we experienced in the original review: although we intended to assess the quality of the primary studies based on the general prompts (see above), this was simply not possible in the majority of the studies. Arguably, it is possible to contact authors with the request to provide this information ad hoc, but this does not guarantee actual obtaining this information. For instance, authors may not respond to such requests, authors may not have the necessary information any more, contact information may have changed, etc. Thus, adequate reporting seems to be a necessary condition for quality assessment (Carroll & Booth, 2015; Dixon-Woods et al., 2004). A second reason is that judging whether or not elements of a study are adequately reported is said to be less difficult than assessing other features (Carroll et al., 2012). As the quality assessment in relation to meta-integration is particularly complicated, the use of a simple and uncomplicated tool that can assess the quality of qualitative, quantitative and mixed methods papers is strongly recommended (Frantzen & Fetters, 2015).

Based on the literature, we derived a simple checklist for quality appraisal (Carroll et al., 2012; Franzel et al., 2013; Hannes, 2011). The checklist contained five items (see Table 1): four previously used by Carroll and colleagues (2012)[[11]](#footnote-11) and one added by ourselves (i.e. description of sample). This last item was added as this is a major indicator to allow evaluation of transferability of qualitative findings (Hannes, 2011). In order to facilitate the appraisal and to ensure that studies were appraised on the same aspects (and thus to enhance the quality of the assessment), each criterion was defined (see Table 1) (Carroll et al., 2012). Per study, these five reporting criteria were assessed.

Table 1: Reporting criteria checklist

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Categorization** | **Description** |
| Research questions/objectives | Y | The research questions and/or the study objectives are clearly mentioned |
|  | N | The research questions and/or the study objectives are not specified |
| Sampling | Y | The selection of participants is described explicitly (e.g. purposeful, convenience) |
|  | N | The selection of participants is not described explicitly |
| Description of sample | Y | A full description of the characteristics of the sample is given (e.g. demographic characteristics, number of participants) |
|  | N | A full description of the characteristics of the sample is not given (e.g. just mentioning interviews with key informants) |
| Methods of data collection | Y | Details of the data collection method are given (e.g. piloting, topic guides for interviews) |
|  | N | Details of the data collection method are not specified (e.g. just mentioning interviews) |
| Methods of data analysis | Y | Details of data analysis are presented (e.g. recording, transcribing, coding) |
|  | N | Details of data analysis are not presented (e.g. just mentioning data were analysed) |

The second author read the original publications and extracted text fragments that addressed the quality assessment questions (if available) into a checklist form (Carroll et al., 2012). These fragments were retrieved from the introduction and methods sections. The author assigned a yes, no or unclear for each criterion per study. Studies that were assigned a clear yes against three or more criteria were categorized as adequately reported studies. If studies were assigned less than three yes responses, they were categorized as inadequately reported. When the author found it unclear whether or not to categorize a study as adequately reported – which occurred in five publications – the first author independently assessed the reporting criteria of the studies (see Papaiaonnou, 2012, 116). Based on a comparison of both appraisals, the five studies could be categorized. From the 31 primary studies included in our original synthesis, 18 were excluded as they were categorized as inadequately reported studies. The remaining 13 studies were included in the subsequent thematic synthesis as they were categorized as adequately reported studies.

In line with Carroll et al. (2012), we assessed whether the thematic synthesis was affected by the exclusion of the inadequately reported studies. More specifically, we assessed if all themes generated by the original synthesis were also generated by the synthesis after excluding inadequately reported studies. Subsequently, we evaluated the extent to which the themes retained their original thickness. This means that we studied if the new themes included all relevant information and alternative perspectives (Carroll et al., 2012). This post hoc assessment of the impact of exclusion on the results of the synthesis was possible because we recorded evidence for each theme per study (Carroll & Booth, 2015).

As in our original scoping review, we conducted a basic convergent meta-integration and used a thematic synthesis to generate descriptive themes. Our analysis illustrates that excluding 18 studies did not affect the number and nature of retrieved themes: the synthesis of all 13 studies led to the construction of the same themes as found in the original synthesis (see Table 2). This means that not one theme was completely dependent on the findings from the excluded primary studies.

Table 2: Results sensitivity analysis in terms of identification themes

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Original review** | **Sensitivity analysis** | **Difference** |
| **Themes** | **Support (N)** | **Support (N)** |  |
| Lack cooperation | 13 | 4 | 9 |
| Regulatory frameworks, training, professionalization & (mis)conduct | 16 | 9 | 7 |
| Conflicting expectations & extending responsibilities | 3 | 2 | 1 |
| Influence on citizens' rights & accountability | 8 | 4 | 4 |
| Negative consequences PP | 5 | 1 | 4 |
| Conflicts of interest | 3 | 1 | 2 |
| Mission creep - delineation role | 6 | 5 | 1 |
| Disagreement responsibility policing | 1 | 1 | 0 |
| Filling a void | 6 | 2 | 4 |
| More police capacity for core tasks | 7 | 5 | 2 |
| Assistance in police core tasks | 4 | 2 | 1 |

Apart from having an influence on the construction of themes (which was not the case in our sensitivity analysis), the exclusion of inadequately reported studies may affect the thickness or richness of the themes. A first indicator is the number of themes adequately and inadequately reported studies contribute to (Carroll et al., 2012). The contribution of inadequately reported studies to the themes varied from none (one study) to four (two studies). Five inadequately reported studies contributed to one theme, four studies contributed to two themes and five studies contributed to three themes. Expressed in percentages, we see that approximately 11% of the studies contribute to zero themes, 28% contribute to one theme, 22% to two themes, 28% to three themes and 11% contribute to four themes (see Table 3).

Table 3: Comparative contribution to themes

|  |  |
| --- | --- |
| **18 inadequately reported studies** | **13 adequately reported studies** |
| Contribution to  | N | % | Contribution to | N | % |
| 0 themes | 2 | 11,1 | 0 themes | 0 | 0 |
| 1 theme | 5 | 27,8 | 1 theme | 3 | 23,1 |
| 2 themes | 4 | 22,2 | 2 themes | 2 | 15,4 |
| 3 themes | 5 | 27,8 | 3 themes | 4 | 30,8 |
| 4 themes | 2 | 11,1 | 4 themes | 2 | 15,4 |
| 5 themes | 0 | 0 | 5 themes | 2 | 15,4 |

When looking at the adequately reported studies, we see that approximately 23% of the studies contributed to one theme, 15% contributed to two themes, 31% contributed to three themes, 15% contributed to four themes and, finally, another 15% contributed to five themes. Based on this comparison, we see that exclusion of inadequately reported studies does not negatively affect the thickness or richness of the results. Of course, this is only a quantitative indicator.

A second indicator is the amount of nuance or dissonance that may be lost by excluding inadequately reported studies. Our sensitivity analysis illustrated that this was the case for the theme ‘lack of cooperation’. All four adequately reported studies that contributed to this theme stressed a negative relation between (private and public) policing agents: cooperation did not occur or cooperation was hampered by lack of trust or by practical issues. Positive aspects about cooperation between policing agents were not reported in these studies. However, this was reported in an excluded study (Schneider, 2006) in which was found that public police and forensic accounting organisations have a high level of formal and informal cooperation, which mainly resulted from the specialized skills and expertise of forensic accountants, their ability to make a significant contribution to financial crime investigations and their credibility as witnesses in court. This study thus contradicts other studies pointing to difficult or absent cooperation between policing agents. Excluding this study could thus lead to wrong conclusions regarding this theme.

1. **Discussion**

The main contribution of this paper is to provide insights into the assessment of the impact of excluding inadequately reported studies on a meta-integration, using thematic analysis as synthesis method. We had established that the potential effect of critical appraisals in reviews had not been studied extensively. As such, it is not entirely clear which advantages critically appraising actually entails. To this end, two indicators have been used: (1) loss of themes and (2) thickness of themes (richness of detail within the themes and presence of alternative viewpoints or dissonance). We have studied these effects in the context of a scoping review, combining both English and Dutch literature on plural policing, taking into account studies using multiple methods. In this discussion, we first want to go into some points of debate that we were confronted with during this study. These are divided in two major points: quality assessment and basic convergent meta-integration using thematic synthesis.

*Quality assessment: difficulties and how we dealt with them*

In our study, the sensitivity analysis shows that the exclusion of studies after critical appraisal does not impact the nature and number of retrieved themes. It did, however, impact the level of nuance that can be provided in analysing the remains studies. This is an important remark to make, as this might lead to wrongful interpretations of research results. We have looked for potential explanations for the impact of the sensitivity analysis. First of all, we found much *indirect* information on the research questions. Only a small proportion of the literature in our sample directly addressed the research questions of our study. Based on this finding, it would not be surprising that excluding (more than) half of the studies will lead to less support for the themes (i.e. less studies per theme). This however, was not the case. The indirectness of the information, potentially leading to a lower confidence in the evidence of the literature, is difficult to correct for, especially when this is partly due to the original research question. However, it does lead us to the recommendation that before carrying out a scoping review, a more low-level review (such as a narrative review) might be useful to work on the research questions (to make them apt for the scoping review). The nature of the review (systematic review versus scoping review) probably also impacts the degree of critical appraisal that should be used. In a scoping review, the aim is to map existing literature on the basis of a more broad question, with a view on a narrative integration of evidence, while in a systematic review, on the basis of a more narrow question, a qualitative synthesis is strived for (Peters et al, 2015). The latter also implies a quality appraisal, while a scoping review, because of its aim ‘to scope’ might afford to be less critical. This after all avoids the danger of losing information and nuance in the review.

When considering the way in which the critical appraisal is carried out, we found it difficult to provide clear-cut judgements with regard to reporting criteria. This in spite of the predefined criteria that we used. We give a few examples. For instance: when can you decide if the sampling strategy is explicitly described? Often we found that the sampling strategy was not explicitly named, although it was possible to infer the strategy based on the description. In that case, we assigned an ‘unclear’. Similarly, how to decide if sufficient details regarding the methods of data collection are provided? We sometimes came across studies that mentioned the duration of interviews and the topics covered in the interviews but provided no other details. How to decide whether or not this is sufficient?

We see two solutions for this: (1) carefully registering all text fragments per study in a checklist form so that comparison between studies is possible: if one assigns a yes in that study based on those text fragments, it becomes easier to decide whether or not to assign a yes in another study; (2) independent assessment by two or more reviewers and discussion to validate the decisions. Nevertheless, when compared to other structured appraisal instruments - which are reported to result in limited consensus (Noyes et al., 2008) - assessment of reporting criteria might still be more easy (Carroll et al., 2012). It does, after all, mainly imply a very close reading of the papers. This method hence seems especially attractive for novice reviewers of reviewers without extensive knowledge of evaluation criteria for qualitative or quantitative studies (Hannes, 2011).

A second difficulty involves the balance between the *quality of reporting and the significance or relevance of results*: sometimes a study is excluded based on the inadequate reporting, although the results presented in the study are of high relevance for the research question. Handbooks refer to seeking those studies that are central to your research question. But what should be done if it becomes clear that these central studies will not pass the quality appraisal phase? We would opt for keeping these studies in the review, but on the basis of which criterion (apart from the content) can we do this?

Thirdly, we wonder about the balance between *quality of reporting and quality of study*: together with Dixon-Woods et al. (2004) and Hannes et al. (2010), we acknowledge the possibility that inadequately reported studies may have been well conducted. Nevertheless, in the absence of adequate reporting, this can only be a hunch, a gut feeling, an instinct which is not sufficient from a scientific point of view. We could of course recommend to look for primary materials in which the method has been discussed extensively, but it is debatable how much effort should be gone through in looking for these?

Of course, given the word limit imposed by many academic journals, authors are compelled to carefully select what they write. Although it is fully understandable that researchers have their own preferences as to what they focus on (e.g. empirical results, theoretical framing of results) and anonymous reviewers may request limited methodological information, we would nevertheless stress the need to offer sufficient (methodological) information in order to be able to assess the quality of the research. Limited methodological information has been reported previously (Carroll & Booth, 2015; [Major & Savin-Baden, 2011](#_ENREF_20)) and unfortunately it appears to remain an issue. Deficient reporting has previously led some academics to propose standards of reporting in qualitative research (e.g. Tong, Sainsbury & Craig, 2007 in Carroll & Booth, 2015). We align with this suggestion and urge authors to adequately report minimum information on methodology.

*Basic convergent meta-integration and thematic synthesis: discussion*

We experienced that it was impossible to use completely reproducible and transparent search and selection strategies and to construct an inherently reproducible synthesis ([Dixon-Woods et al., 2006)](#_ENREF_7). For instance, the formulation of the first research question embodies a subjectivity very early on in the review: a ‘danger’ is not clearly defined, as a result of which a subjective assessment takes place right at the beginning of the review. When selecting studies, we had to assess whether or not the publication would address potential dangers of plural policing. The second research question yielded very little empirical material. Even less findings related to the third research question, which was dropped early on in the review. The choice for including studies in two languages was founded on the conviction that many police researchers from the Netherlands and Belgium do not publish in English. We therefore aimed to include these studies in our review. This undoubtedly has resulted in a bias in this review, which we have taking into account in advance. To assess the impact on the thematic synthesis and comprehensiveness, we checked our line by line coding of the publications in Dutch. These codings fitted the English descriptive themes. Of course, this has not been tested, so future research should perhaps test this more extensively before final recommendations can be made to use this method (sequential coding; line-by-line coding in different languages).

We suggest that basic convergent meta-integration and thematic synthesis are useful methods to integrate qualitative and qualitative research findings in criminology. Regarding the basic convergent meta-integration as proposed by Frantzen and Fetters (2015), we experienced no need for intra method synthesis. Failing to synthesize the findings from the qualitative and transformed papers separately was not experienced as a problem and we do not immediately see how this would be necessary. Whether or not conducting an intra method synthesis would have led to different results has however not been tested. Future research could analyse the usefulness of such an intra method synthesis by comparing end results obtained with and without intra method synthesis in order to improve our knowledge on basic convergent meta-integration.

The fact that we limited ourselves to the first two steps of thematic synthesis was not based on an impossibility to conduct the third step, but on the aim of our review and the nature of the topic. We clearly chose to conduct an integrative more than interpretive review. Whereas the first is more focused on summarizing data, the latter is more directed towards the development of concepts and theories that integrate those concepts ([Dixon-Woods et al., 2005](#_ENREF_6)). In these two steps, we did encounter some difficulties. However, we suspect that these did not arise from the method in se, but more from the limited information found per research question. We ended up with thirty-one publications that only offered indirect information on our research questions. During the line-by-line coding, not much text was found to substantiate each free code. We had many free codes, but only limited text per free code. For us, this was a stunning finding as contemplative publications offer quite a lot of attention to the items covered in our review. Apparently, empirical research is lagging behind. The limited information on the research questions and the highly varying research settings and contexts complicated finding relevant similarities, differences and other relations, which form the basis of the descriptive themes. We tried to correct for the lack of direct information by informing the readers about this ‘indirectedness’ in the original review. In that way, we hoped to make clear to readers that the themes were based on indirect information and that therefore carefulness is needed in interpreting or using the results of the review. Thematic synthesis is developed to integrate findings from multiple qualitative studies ([Thomas & Harden, 2008](#_ENREF_33)) but we have found that it also enables to integrate findings from quantitative or mixed methods designs. This is done by coding textual fragments, derived from the numerical data. As such, we did not use the presented statistical data, but used the textual explanations. Therefore, thematic synthesis does not seem suitable to integrate purely quantitative data to estimate effect sizes and/or generate numerical outcomes. If reviewers are satisfied with only using textual fragments of quantitative studies, our experience is that it does not pose a problem to use thematic synthesis to integrate findings from qualitative and quantitative studies. We must admit that the studies, based on a quantitative research, did provide sufficient useful textual fragments which may influence our conclusion that quantitative evidence can be analysed with thematic synthesis. Other quantitative research, with less textual results and more numerical data, may be more difficult to summarize with thematic synthesis.

1. **Conclusion**

Can we make a strong case or not with regard to excluding inadequately reported studies in scoping reviews? The answer is both yes and no.

Our analysis suggests that the excluded papers had no impact on the identification of descriptive themes. Nevertheless: they did have an impact on thickness of themes and more specifically on the nuances and dissonance. Given the aim of a scoping review (mapping the literature on a topic) it seems relevant to include all viewpoints on a topic, including alternative viewpoints and nuanced visions. Precisely the aim of a scoping review lies in getting a good view on what is already known in the literature. Therefore, excluding alternative viewpoints, dissonance or nuances might be pernicious. This leads to the conclusion that, to obtain a fuller account of the phenomenon under scrutiny, it seems beneficial to include inadequately reported studies. This is perhaps not in line with previous research (Carroll et al., 2012; Franzel et al., 2013, Noyes & Popay, 2007; Thomas & Harden, 2008), but their conclusions relate to systematic reviews, not scoping reviews. Of course, including inadequately reported studies in a scoping review does not imply ‘not doing anything with the appraisal. Reviewers should be aware of the quality of reporting and should at least inform their readers about the inclusion of lower quality studies in the review. As such, responsibility for what is done with that information, lies in hands of the informed readers.

**References**

Arksey, H. & O’Malley, L. (2005). Scoping studies: towards a methodological framework, *International Journal of Social Research Methodology: Theory & Practice, Vol.8*(1), 19-32.

Baarda, D. B., De Goede, M. P. M., & Teunissen, J. (2009)*. Basisboek kwalitatief onderzoek: handleiding voor het opzetten en uitvoeren van kwalitatief onderzoek*. Groningen: Stenfert Kroese.

Barnett-Page, E. & Thomas, J. (2009). Methods for the synthesis of qualitative research: a critical review, *BMC Medical Research Methodology, Vol. 9* (59),1-11.

Boeije, H.R., van Wesel, F., & Alisic, E. (2011). “Making a difference: towards a method for weighing the evidence in a qualitative synthesis”, *Journal of Evaluation in Clinical Practice, Vol. 17(4),* 657-663.

Boels, D. and Verhage, A. (2016). “Plural policing: a state-of-the-art review”, *Policing: An International Journal of Police Strategies and Management, Vol. 39(1)*, 1-16.

Booth, A. (2012). “Systematic approaches to the literature”, in Booth, A., Papaioannou, D. & Sutton, A. (Eds.), *Systematic approaches to a successful literature review* (pp.17-35), London: Sage.

Booth, A. (2001). *Cochrane or cock-eyed? How should we conduct systematic reviews of qualitative research?* Paper presented at the Qualitative Evidence-based Practice Conference. Taking a critical stance, Coventry University.

Carroll, C. & Booth, A. (2015). Quality assessment of qualitative evidence for systematic review and synthesis: is it meaningful, and if so, how should it be performed? *Research Synthesis Methods, Vol. 6*(2), 149-154.

Carroll, C., Booth, A. & Lloyd-Jones, M. (2012). "Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews", *Qualitative Health Research, Vol. 22(10),* 1425-1434.

Cruzes, D. S., & Dyba, T. (2011). *Recommended steps for thematic synthesis in software engineering*. Paper presented at the International Symposium on Empirical Software Engineering and Measurement, Banff.

Daudt, H., van Mossel, C., & Scott, S. (2013). "Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework", *Medical Research Methodology, Vol. 13(48),* 1-9.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Services Research & Policy, 10*(1), 45-53.

Dixon-Woods, M., Bonas, S., Booth, A., Jones, D. R., Miller, T., Sutton, A. J., . . . Young, B. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research, 6*(1), 27-44.

Dixon-Woods, M., Booth, A., & Sutton, A. J. (2007). Synthesizing qualitative research: a review of published reports. *Qualitative Research, 7*(3), 375-422.

Dixon-Woods, M., Fitzpatrick, R., & Roberts, K. (2001). Including qualitative research in systematic reviews: opportunities and problems. *Journal of Evaluation in Clinical Practice, 7*(2), 125-133.

Dixon-Woods, M., Shaw, R. L., Agarwal, S., & Smith, J. A. (2004). The problem of appraising qualitative research. *Quality and Safety in Health Care, 13*, 223-225. doi: 10.1136/qshc.2003.008714

Finfgeld-Connett, D. (2013). Use of content analysis to conduct knowledge-building and theory-generating qualitative systematic reviews. *Qualitative Research, 14*(3), 341-352.

Finfgeld-Connett, D. (2014). Metasynthesis findings: potential versus reality. *Qualitative Health Research, 24*(11), 1581-1591.

Frantzen, K.K., & Fetters, M.D. (2015). "Meta-integration for synthesizing data in a systematic mixed studies review: insights from research on autism spectrum disorder", *Quality and Quantity,* Published online 30 August 2015. DOI 10.1007/s11135-015-0261-6.

Franzel, B., Schwiegershausen, M., Heusser, P. & Berger, B. (2013). "How to locate and appraise qualitative research in complementary and alternative medicine", *BMC Complementary & Alternative Medicine, Vol. 13(125),* 1-10.

Green, S., Higgins, J. P. T., Alderson, P., Clarke, M., Mulrow, C. D., & Oxman, A. D. (2008). Introduction. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 3-9). Chichesters: Wiley.

Hammerstrom, K., Wade, A., & Klint Jorgensen, A.-M. (2010). Searching for studies: A guide to information retrieval for Campbell Systematic Reviews (pp. 74). Oslo: The Campbell Collaboration.

Hannes, K. (2011). "Chapter 4: Criticial Appraisal of qualitative research"*,* in Noyes, J., Booth, A., Hannes, K., Harden, A., Harris, J., Lewin, S., & Lockwood, C. (eds), *Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions*, Version 1, Cochrane Collaboration Qualitative Methods Group.

Hannes, K., & Claes, L. (2007). Learn to Read and Write Systematic Reviews: The Belgian Campbell Group. *Research on Social Work Practice, 17*(6), 748-753. doi: 10.1177/1049731507303106

Hannes, K. & Harden, A. (2011). "Multi-context versus context-specific qualitative evidence syntheses: combining the best of both", *Research Synthesis Methods, Vol. 2(4),* 271-278.

Hannes, K., & Lockwood, C. (Eds.). (2012). *Synthesizing qualitative research. Choosing the right approach*. West Sussex: John Wiley & Sons, Ltd.

Hannes, K., Lockwood, C., & Pearson, A. (2010). A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. *Qualitative Health Research, 12*, 1736-1743. doi: 10.1177/1049732310378656

Hannes, K., & Macaitis, K. (2012). A move to more systematic and transparent appraoches in qualitative evidence synthesis: update on a review of published papers. *Qualitative Research, 12*(4), 402-442.

Lee, R. P., Hart, R. I., Watson, R. M., & Rapley, T. (2015). Qualitative synthesis in practice: some pragmatics of meta-ethnography. *Qualitative Research, 15*(3), 334-350.

Major, C. H., & Savin-Baden, M. (2011). Integration of qualitative evidence: towards construction of academic knowledge in social science and professional fields. *Qualitative Research, 11*(6), 645-663.

Manning, N. (2012). Conclusion. In K. Hannes & C. Lockwood (Eds.), *Synthesizing qualitative research. Choosing the right approach* (pp. 161-171). West Sussex: John Wiley & Sons, Ltd.

Noyes, J. & Popay, J. (2007). "Directly observed therapy and tuberculosis: how can a systematic review of qualitative research contribute to improving services? A qualitative meta-synthesis", *Journal of Advanced Nursing, Vol. 57(3),* 227-243.

Noyes, J., Popay, J., Pearson, A., Hannes, K., & Booth, A. (2008). Chapter 20: Qualitative research and Cochrane reviews. In J. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*: The Cochrane Collaboration.

Papaioannou, D. (2012). "Assessing the evidence base", in Booth, A., Papaioannou, D. & Sutton, A. (Eds.), *Systematic approaches to a successful literature review* (pp.97-124), London: Sage.

Paterson, B. L. (2012). "It looks great but how do I know it fits?": an introduction to meta-synthesis research. In K. Hannes & C. Lockwood (Eds.), *Synthesizing qualitative research. Choosing the right approach* (pp. 1-20). West Sussex: John Wiley & Sons, Ltd.

Pearson, A. (2004). Balancing the evidence: incorporating the synthesis of qualitative data into systematic reviews. *JBI Reports, 2*, 45-64.

Peters, M. D., Godfrey, C. M., Khalil, H., McInerney, P., Parker, D., & Soares, C. B. (2015). "Guidance for conducting systematic scoping reviews". *International journal of evidence-based healthcare, 13(3),* 141-146.

Pham, M.T., Rajic, A., Greig, J.D., Sargeant, J.M., Papadopoulos, A. & McEwen, S.A. (2014). "A scoping review of scoping reviews: advancing the approach and enhancing the consistency", *Research Synthesis Methods, Vol. 5(4)*, 371-385.

Popay, J., Rogers, A., & Williams, G. (1998). Rationale and Standards for the Systematic Review of Qualitative Literature in Health Services Research. *Qualitative Health Research, 8*(3), 341-351. doi: 10.1177/104973239800800305

Ring, N., Ritchie, K., Mandava, L., & Jepson, R. (2010). A guide to synthesising qualitative research for researchers undertaking health technology assessments and systematic reviews: NHS Quality Improvement Scotland, University of Stirling, School of Nursing, Midwifery and Health.

Schneider, S. (2006). "Privatizing economic crime enforcement: exploring the role of private sector investigative agencies in combating money laundering", *Policing & Society, Vol. 16(3),* 285-312.

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). Quality in Qualitative Evaluation: a framework for assessing research evidence. A quality framework. (pp. 18). London: Government Chief Social Researcher's Office. Cabinet Office.

Terpstra, J., & Stokkom, B. v. (2015). Plural Policing in Comparative Perspective. Four Models of Regulation. *European Journal of Policing Studies, 2*(3), 325-343.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical research Methodology, 8*(45), 1-10.

Toye, F., Seers, K., Allcock, N., Briggs, M., Carr, E., Andrews, J. & Barker, K. (2013). "'Trying to pin down jelly' - exploring intuitive processes in quality assessment for meta-ethnography", *Medical Research Methodology, Vol. 13(46),* 1-12.

Verhage, A. & Boels, D. (2015). Systematische reviews van kwalitatief onderzoek binnen de criminologie, *Panopticon, Vol 36 (3),* 302-311.

1. Currently policing is not solely executed by the public police anymore. Instead, multiple actors (e.g. private police, inspectorate services, citizens) are engaged in policing activities (e.g. order maintenance, law enforcement, crime investigation). The term plural policing refers to the fact that policing services are delivered in the public, semi-public and private domain by a complex of public and private bodies and agencies (for results of the review, see Boels and Verhage, 2016). The aim of our scoping review was to identify the literature on the dangers of plural policing and on the impact of plural policing on core police tasks. [↑](#footnote-ref-1)
2. Some of these have been thoroughly explained such as content analysis ([Finfgeld-Connett, 2013](#_ENREF_11)), meta-ethnography ([Lee, Hart, Watson, & Rapley, 2015](#_ENREF_19)) and thematic synthesis (Cruzes & Dyba, 2011). [↑](#footnote-ref-2)
3. For us, empirical research implies that the publication needed to refer to newly collected data. Empirical was defined as “*established through observation what is happening in reality*” (Baarda, De Goede & Theunissen, 2009). This implies data gathering, through either qualitative, or quantitative methods, or a combination of both. [↑](#footnote-ref-3)
4. Blurring AND boundaries OR lines; extended policing family; new security complex; policing AND commodification; policing AND diversification; policing AND fragment\*; policing AND hybrid; policing AND network; policing AND third party; polic\* AND plural; polic\* AND complex; polic\* AND function. [↑](#footnote-ref-4)
5. Toezicht; handhaving; netwerkstructuur; politie AND plural\*; plural AND policing; politiefunctie; veiligheid AND fragmentering; veiligheidscomplex; veiligheidshandhaving; vervagende AND grenzen. We chose to include English key terms to identify studies in Dutch, as we found that these terms are sometimes also used in Dutch publications. [↑](#footnote-ref-5)
6. In addition, we registered the dates of search, time span of literature (in some databases, such as the Campbell Library, studies were not accessible as from 1990, but only for instance as from 2003), places of keywords search (title, abstract full text) plus reason why not searched in title and extra search features (if applicable). [↑](#footnote-ref-6)
7. Three publications were based on the same empirical data, so in fact two studies used mixed methods. [↑](#footnote-ref-7)
8. For examples we refer to [Dixon-Woods et al. (2004)](#_ENREF_10), [Hannes, et al. (2010](#_ENREF_17)), [Popay, et al. (1998](#_ENREF_26)), Spencer, [et al. (2003](#_ENREF_30)). [↑](#footnote-ref-8)
9. 12 Meaning that the methods used are mentioned by the authors in the introduction and/or methods-section. [↑](#footnote-ref-9)
10. Whereas one study for instance focused on the evaluation of partnerships between the police and local groups engaged in policing in fragile African states, another one focused on the impact of civilian policing on the legitimacy of the public police, the extent of public support for civilian policing and their impact upon crime reduction in the UK. [↑](#footnote-ref-10)
11. These four criteria are included in different appraisal checklists and tools, such as the Critical Appraisal Skills Programme (CASP) and the Evaluation Tool for Qualitative Studies (ETQS) (Carroll et al., 2012). [↑](#footnote-ref-11)