Writing a literature review
Methodological considerations
Goddiksen, Mads; Nielsen, Søren Saxmose; Sandøe, Peter

Publication date:
2017

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Writing a literature review
- Methodological considerations

Mads Goddiksen, Søren Saxmose Nielsen & Peter Sandøe
This is an unpublished manuscript for a teaching note on methodological issues related to literature reviews. The material was written in the fall of 2017 by:

Mads Goddiksen¹, Søren Saxmose Nielsen² & Peter Sandøe¹².

¹Department of Food and Resource Economics, University of Copenhagen
²Department of Veterinary and Animal Sciences, University of Copenhagen

The material can be cited as: Goddiksen, M., Nielsen, S.S. & Sandøe, P.: Writing a literature review: methodological considerations (unpublished manuscript Dec. 2017)
Preface

This note was developed for the bachelor courses in philosophy of science for veterinary and animal science students at University of Copenhagen. It covers a theme that has generally been neglected in standard philosophy of science and philosophy of science teaching: The function of literature reviews in scientific methodology and the methodological challenges related to assessing and discussing conflicting evidence in a literature review. We have chosen to focus on this topic because we believe it is central to the process of developing reliable scientific results, and because it brings together central themes in philosophy of science including causality and scientific uncertainty. At the same time the topic is of immediate relevance to the students who often do project work as a part of the philosophy of science course or begin writing their bachelor project immediately after the course.

Because the text is written as a supplement to the material used in the courses of interest, noticeably Videnskabsteori for de biologiske fag (Andersen et al. 2006), we have chosen to mainly refer to this rather than relevant primary literature.

Acknowledgements

In the development of this note we have received valuable input from Trine Dich and other members of the philosophy of science teachers group at the University of Copenhagen.
Contents

Preface ........................................................................................................................................... 1
Acknowledgements ........................................................................................................................... 1

1 Introduction .................................................................................................................................. 3

1.1 Structure and intended learning outcomes ............................................................................. 4

2 Why make a literature review? .................................................................................................. 5

2.1 Hypothesis testing reviews: an example .................................................................................. 5

2.2 Methodological requirements for different types of reviews .................................................. 7
Section 2 reflection questions ......................................................................................................... 8

3 Finding relevant sources ............................................................................................................. 8
Section 3 Reflection questions ....................................................................................................... 10

4 Reading and evaluating your sources .......................................................................................... 11

4.1 Looking for uncertainties ........................................................................................................ 11

4.1.1 Material error ..................................................................................................................... 11

4.1.2 Observational error ........................................................................................................... 14

4.1.3 Conceptual error ............................................................................................................... 14

4.2 Placing the source in the Pyramid of Evidence ...................................................................... 15

4.2.1 Limitations of the Pyramid of Evidence .......................................................................... 19

4.3 Publication bias ....................................................................................................................... 20
Section 4 Reflection questions ....................................................................................................... 22

5 Synthesizing your sources ......................................................................................................... 22

5.1 Grouping your sources: an example ....................................................................................... 23

5.2 Considering uncertainties ........................................................................................................ 25
Section 5 reflection questions ....................................................................................................... 26

6 Summary: The essential steps in making a literature review .................................................... 27

7 Perspectives .................................................................................................................................. 29
Section 7 reflection questions ....................................................................................................... 30

References ........................................................................................................................................ 30
1 Introduction

Think of a typical picture of a scientist. Often such a picture will show a white male in a white lab coat working alone in a laboratory (cf. Fig. 1). Pictures like these misrepresent modern science in a number of ways (Allchin 2013, Ch. 3). One of these is that they convey the common idea that scientists spend most of their time in the laboratory doing experiments and gathering data. For sure, experiments and data collection are important, even essential, parts of natural science, but there are several other steps in the process of getting to a reliable scientific result that can be just as important, difficult and time consuming as laboratory work.

One of these steps is finding out what “we” – the scientific community – already know about a given topic. This is known as making a literature review. To make literature reviews is an essential part of doing science. Pick almost any article published in a scientific journal, and you will find that the introduction contains a literature review providing an overview of the existing research on the topic considered in the article, and a brief discussion of what we already know, and where our knowledge is lacking. Such introductory reviews can be difficult to make because the existing literature is so extensive that it can be very difficult to get an overview. In such cases it can be useful to draw on dedicated review articles i.e. articles that consist solely of a literature review, often aiming to provide other researchers with an overview of a given research topic.

Because literature reviews are central to any scientific project, it is also important that students learn the art of writing a literature review. For many students, this process starts when they begin work on their first major project and accelerates when writing the bachelor project, which in many cases consists solely of a literature review.

If you have not yet tried to do a detailed literature review, you might think: how hard can it be? You just have to collect the relevant sources, read them, give a summary of what they say and end by stating the conclusions that can be drawn from the sum of the sources. This is basically what you have to do, but there can be a number of challenges related to this effort. For one, it is very likely that you will find several
hundred research articles on a given topic. How do you sort through all of these without having to read them all? Furthermore, it is not uncommon to find sources that reach seemingly very different conclusions. Here it will firstly be important to find out whether there is genuine disagreement or whether the sources really answer different research questions. In cases where the sources answer the same research question we say that there is conflicting evidence on what the “correct” answer to the question is. It is not very satisfactory, in such cases, to simply conclude in your literature review that one source says this and one says the other. The good literature review should also include a discussion of what can be learned from comparing the available sources with each other. To engage in such a discussion, the reviewer, must among other things, be able to critically evaluate the methods used in the different studies, the quality of the data obtained, and the relevance of the detailed research questions asked. When doing so, questions about different types of scientific uncertainty become central. There are thus a number of important and interesting methodological challenges related to making a good literature review.

1.1 Structure and intended learning outcomes

In this note we will discuss some of the central steps in making a literature review to be included e.g. in a bachelor project. Once you have read this note and attended the classes on the topic, you should be able to:

Knowledge:

✓ Explain the different functions of literature reviews in scientific methodology
✓ Explain the most important steps in making a literature review
✓ Explain the difference between the different types of studies described in the Pyramid of Evidence (PoE).

Skills:

✓ Place individual studies in the PoE and assess the importance of this categorization in relation to a given research question.
✓ Group sources into relevant categories for use in a review on a given research question.

Competences:

✓ Discuss the role of literature reviews in developing reliable scientific results
✓ Discuss the internal and external validity of a given study in relation to a given research question

To achieve these aims this note proceeds as follows: First we discuss the many different purposes that literature reviews serve in science (Sec. 2). We then proceed to discuss the methodological challenges related to making a literature review aiming to determine what the existing scientific literature can tell us about a given research question. The first challenge related to this is identifying relevant sources (Sec. 3). Once
relevant sources have been found, it is time to read and assess the sources individually (Sec 4), before beginning to actually write the review (Sec 5). The whole process is summarized in Sec. 6. We end with a broader look at the role of literature reviews in scientific methodology (Sec. 7).

2 Why make a literature review?
There are many reasons why a researcher or student might want to make a literature review. The most basic motivation is a desire to find out what the existing scientific literature can tell us about a given topic, either because we simply want to know, or because we are looking for areas where we might contribute to the expansion of our existing knowledge. Such reviews are not necessarily intended for publication, but can simply serve as an early explorative step in the research process. Later on, the results of these early efforts can be incorporated into an academic text.

Publications in the natural sciences most commonly aim to present new interesting results obtained through empirical methods. In such publications the aim of the literature review in the introduction is to place the study in a broader context and relate it to previous knowledge, thus making the empirical results interesting to others.

Literature reviews play other central roles in other types of research publications (Randolph 2009). As mentioned in the introduction, some research publications consist solely of a literature review often aiming to provide a fairly complete overview of the research on a given topic, or to show its historical development.

Finally, researchers sometimes make literature reviews to test a hypothesis (see Sec. 2.1).

Readers of literature reviews of course often read literature reviews in order to achieve the same aims as the author had in mind when writing them, but in addition, the reader may also aim to assess whether the author of the review has sufficient overview of the research field in question, understands the basic theory, and knows what the most important variables are. Answering these basic questions can be an important part of the overall assessment of the trustworthiness of other results that may be presented in the article. As a student this can useful to keep in mind, e.g. when writing a review for a bachelor project in order to make sure that the assessors get the answers that they need (for further details see Randolph (2009)).

2.1 Hypothesis testing reviews: an example
Literature reviews can in some cases be used to test a hypothesis. Say you want to test the common hypothesis that conventional free range layer hens (Fig. 2) generally have higher welfare than layer hens kept in cage systems. Testing this hypothesis of course requires empirical data, but there is already an abundance of research on how housing systems affect hen welfare that the hypothesis can be reliably tested without collecting new data.
Welfare is a contested concept, with several definitions, so before embarking on welfare research a definition must be settled on. Subsequently this definition must be operationalized in the form of a set of welfare indicators. Individual studies of hen welfare tend to focus on one or a few welfare indicators, such as mortality or foot health, and compare only two different housing systems, e.g. free range vs. conventional cages. The results of these studies may point in many different directions, and there can even be disagreement among studies that use similar methods. Thus one study (Shini 2003) found that free range hens generally are less stressed than hens in cages, while another study (Mench et al. 1986), found no difference in stress levels when comparing hens in cages with hens in non-cage systems. It is thus not hard to find sources that can be used to support an argument for the claim that free range hens have higher welfare and sources that will not support it. The challenge is to find out which conclusion the sum of the available evidence supports.

Lay and colleagues took up this challenge in 2010 (Lay et al. 2010). They compiled more than 200 studies on hen welfare in different housing systems and compared the results. They concluded that stress is a problem in all housing systems, but hens in so-called furnished cages (the only cages for laying hens currently allowed in the EU) generally have lower mortality and less feet problems than hens in other housing systems. Free range hens, on the other hand, generally have access to more space than caged hens which allows them to better express certain natural behaviors, such as dust bathing and grooming, which is generally associated with higher welfare. However, this advantage compared to furnished cages comes at a price, as free range hens besides the problems already mentioned are more likely to be infected by pests and parasites, to suffer from painful bone fractures, particularly keel bone fractures, which are found in more than half of non-cage hens and last to be more prone to behavioural problems such as feather pecking and cannibalism.

Lay and colleagues thus found that at the time, there was no evidence indicating that free range hens generally had higher welfare than hens in other housing systems.

---

1 For a general discussion of the challenges related to animal welfare assessments see Sandøe & Christiansen (2008), Ch. 3.
2 Here the methodological considerations connect to the discussion on cherry picking in Goddiksen & Johansen (2017). Cherry picking, i.e. deliberately discarding sources because they do not support your own hypothesis, is problematic not only from a methodological point of view, but also from an ethical perspective and may thus be seen as a violation of good scientific practice.
This conclusion was confirmed by a study that came out in 2010 and which was not included in the review by Lay and colleagues\(^3\). This study by Sherwin and colleagues (2010) is one of the most comprehensive studies comparing hen welfare in four different housing systems, including furnished cages and free range. The welfare of the hens was measured using multiple methods focusing on multiple variables, and the welfare of the hens was measured multiple times all the way from placement to slaughter. Based on this extensive dataset, Sherwin and colleagues went as far as to conclude that “the welfare of the birds in the FC [furnished cage] systems was better than that of the birds in the other systems” (p. 498).

The study by Sherwin and colleagues may not be sufficient to reject the claim that free range hens have better welfare than caged hens, as there are some uncertainties associated with it. For instance, four of the six furnished cage flocks that participated in the study were on the same farm and attended by the same personnel. Was it actually the specific people rather than the general housing system that made the difference? This cannot be assessed solely by looking at the data from the study itself. However, by looking at other studies performed on farms from all over the world and seeing that their results point in the same direction tells us that it was most likely the housing system, and not the specific workers that made the significant difference.

Based on the available evidence it is thus possible to conclude that of the conventional housing systems currently in use, the hens in free range systems do not generally have higher welfare than hens in other systems – at least as long as welfare is defined in terms of absence of suffering. Our initial claim therefore, in light of the existing evidence, seems indefensible. The case illustrates one of the many different functions of literature reviews, namely as a tool for checking claims made based on common sense, press reports and the like. It should be added that, of course, new research may modify such conclusions. Thus, even though the evidence provided by Sherwin and colleagues is comprehensive it is, of course not exhaustive. For example, it does not evaluate the positive value to the free range hens of behavioral opportunities.

### 2.2 Methodological requirements for different types of reviews

The purpose of a literature review determines its methodological requirements. Generally, the requirements for reviews that aim to be more or less complete or to test a specific hypothesis are stricter than for other types of reviews. The former should be systematic reviews where the decision on what sources to include in the review is not based on whatever the author has already read or likes, but on author independent, explicitly defined inclusion criteria, that are used to search through all the available literature.

If a review is used to test a hypothesis, researchers sometimes not only perform a systematic review of the literature, but also combine the data from all the individual studies and perform statistical analyses on this

---

\(^3\) The two studies do not mention each other, presumably because Lay and colleagues prepared their review before Sherwin and colleagues published their study 2010, but the review by Lay and colleagues was not published until 2011.
new, much more extensive data set. This is known as a meta-analysis, and is generally considered to be one of the most reliable methods for testing a hypothesis (cf. Sec. 4.2). Reviews that do not contain a meta-analysis are referred to as qualitative reviews even when they review quantitative results (Pautasso 2013).

In this note we are not going to discuss the details of how to make a systematic review or a meta-analysis. Instead we will focus on the types of reviews that bachelor- and master students are typically asked to perform: Qualitative reviews of an extensive but incomplete set of sources aiming to find out what the sources can tell us about a specific research question.

The most important characteristic of such a review is that it goes beyond merely describing the individual sources. The good review should also:

a) provide the reader with an overview of the sources by grouping them e.g. according to whether they support or contradict a specific hypothesis
b) discuss the reliability of the individual sources
c) discuss the conclusions that can be drawn by comparing the sources to each other.

Furthermore, the good review may also:

d) point to areas of significant scientific uncertainty that call for further research.

The first step towards these aims is finding relevant sources.

Section 2 reflection questions
1. What is a systematic review?
2. What is a qualitative review?
3. Have you written a literature reviews before? If so, what were the purposes of these reviews? What were the biggest challenges that you faced when writing them?

3 Finding relevant sources

Finding relevant sources is generally one of the first challenges faced when making a literature review. Often this challenge is linked to the challenge of finding a suitable research question. The research question (perhaps with a few sub-questions) should be precise, interesting and relevant, and at the same time not so general that it is impossible to cover within one study (Dahl et al. 2016, Ch. 5). Some bachelor, master and even Ph.D. students work on predefined projects designed by their supervisors. If not, the initial literature search can give important indications on whether the initial research question is suitable. If the question is too broad, the literature search is likely to yield an overwhelming amount of sources. The initial literature
search may also show that the initial question has either already been answered or otherwise show that it is not actually an interesting question.

Finding the relevant sources for a *complete* literature review raises two challenges: Finding *all* the relevant sources, and finding *only* the relevant sources. Solving one challenge often intensifies the other. If the search is broadened to try and find all the relevant sources, there is a risk of being overwhelmed with material, much of which will be of little relevance. Narrowing the search to try and find only the most relevant material, on the other hand, increases the risk of missing important sources.

For the student making a literature review e.g. for a bachelor project, the task is a bit easier for several reasons: First, the student is writing a different type of review that is not meant to be *complete*, but simply a review of the most relevant sources on the topic in question. Second, the student will have a supervisor who may be able to help identify the most relevant sources. In addition, the university library offers specialized courses in literature search techniques, and the librarians employed there are specialists in seeking relevant information (see also Dahl et al. (2016), Ch. 10). We will therefore not go into details about the specific techniques that can be used to find relevant literature. Instead we focus on what to do once a set of relevant sources has been identified.

Once you have a set of relevant sources, the next step can be to sort them very roughly into those that are sufficiently trustworthy to be included into an academic project, and those that are not. A detailed analysis of whether a source is trustworthy requires careful critical reading and perhaps a checkup on the authors, and the sources referred to in the source. Although you should do this for all the sources that you rely heavily upon in your review (as discussed in Sec. 4), it is probably not feasible to do so for *all* the sources you find through the literature search, simply due to lack of time. You will therefore have to use more coarse measures initially. A good starting point is to discard all the sources that have not been through a thorough quality control prior to publication.

The sources that go through the strictest quality control prior to publication are research articles submitted to well-known academic journals and official reports from major official institutions. Articles submitted to well-known academic journals are not only scrutinized by the journal editor; the manuscripts also go through *peer review*. In the peer review process the quality and novelty of a manuscript submitted to an academic journal is assessed by a number of experts - usually between two and five - before being accepted for publication\(^4\). Because the well-known journals do not have to worry about not being able to fill up their pages, they can afford to reject the manuscripts that they do not consider to be of sufficient relevance and quality. Peer reviewed sources are therefore generally the most trustworthy sources, but the fact that a source has been through peer review does not mean that the information presented in the source is not biased or

\(^4\) The peer review process is discussed further in Goddiksen & Johansen (2017).
otherwise faulty. It simply means that at least two experts have read the text and concluded that it was of
sufficient quality to be published in the specific journal in question. Still, this is more than can be said of
most blog posts, newspaper articles and websites.5

Textbooks do not generally go through peer review, and can be of very varying quality. On the other hand,
textbooks mainly cover theories and results on which there is a wide scientific consensus. Furthermore, the
textbooks that you have acquired for your courses have been approved by your lecturers (hopefully not just
because they wrote them themselves), and have in this sense been through an additional layer of quality
control before they got to you. Textbooks can thus to some extent be relied upon in your review, particularly
for information on basic theory.

In addition to sorting your sources after trustworthiness, you may also want to limit you set of sources in
other ways, e.g. by focusing only on recent sources. If you sort your sources according to age, you of course
risk discarding relevant sources of high quality just because they are above a certain age, particularly if the
age limit set arbitrarily. To limit the risk, it is helpful, if possible, to supplement you set of recent sources
with a review article covering older studies. This can also help you determine what your age limit should be.
If you have a good review covering studies performed before say 2011, you can include that in your set of
sources, and otherwise focus on sources published after 2011.

Further strategies for limiting your set of sources are discussed in Dahl et al. (2016), Ch. 10.

If you chose to exclude peer reviewed sources from your review based on e.g. their age, it is important that
you state clearly in your review that you have done so, and what exclusion criteria you used.

Once your sources have been roughly sorted, the task is to determine what the core sources can tell you
about the research question that you are interested in. To this end you will of course first have to read
through the individual sources and consider what each of them can tell you on its own.

Section 3 Reflection questions
1. How can you design your literature search to increase the chance of finding sources that both confirm
and reject your hypothesis?
2. Would you include articles relevant to your research question in your review if they were published in
the following outlets: EMBO reports, Scientific American, SEGES publications
(http://svineproduktion.dk/Publikationer). Justify your answer.

5 How do you whether a source has been through peer review? Research articles do not generally come with a
disclaimer on the front page saying “this article has been trough peer review”. However, almost all academic journals
use peer review, although the quality of it varies depending on the journal. So articles and letters (but not necessarily
torialis and opinion pieces) from these journals can therefore safely be assumed to have been through peer review.
For monographs it can be much harder to discern whether they have been through peer review, but often peer
reviewed monographs are part of a more general series described on the publishers web-site.
3. Do you know how to check if a journal uses peer review?
4. Do you know where you can seek advice on how to find relevant literature?

**4 Reading and evaluating your sources**

When reading the individual sources for your review, it is important to make careful notes for each source on e.g. the central messages of the source, the methods used, and the main strengths and weaknesses of the research presented in the source. These notes can then serve as a starting point for the manuscript for the actual review (Pautasso 2013). An important thing to note for each individual source is how strong the arguments presented in the source are. This is known as the *internal assessment* of the source (Hegelund 2000). When making the internal assessment it is important to consider whether there are substantial *scientific uncertainties* surrounding the study presented in the source that could lead you to question the reliability of the conclusions drawn.

**4.1 Looking for uncertainties**

Scientific uncertainty stems from potential errors made in the study itself or in the communication of the results. There are several different types of scientific errors and thus several different types of uncertainties.

It is for instance common to distinguish between *empirical* and *theoretical* uncertainty (Andersen et al. 2006, Ch. 10). Empirical uncertainty is present whenever there is lack of data of sufficient quality. This lack of high quality data can be due to different types of errors in the study. In his typology of scientific error, Allchin (2001) distinguishes between two general types of error that can lead to empirical uncertainty: material errors and observational errors. Observational errors affect our ability to reliably *detect* the phenomenon we are studying. Material errors affect our ability of reliably *generate* the phenomenon we want to study.

**4.1.1 Material error**

Standard examples of material errors include contaminated samples and improper use of laboratory equipment, but sometimes material errors are more subtle (cf. Fig. 3). Many experiments for instance involve dividing a sample population into an “intervention” group and a “control” group. The two groups are assumed to be identical “on average” on all relevant parameters, except for the specific factor that is manipulated in the intervention. This is known as a *controlled trial*. However, if the allocation into control and intervention groups is not done properly (using random allocation), then the two groups will not be identical, even “on average”, which means that it becomes uncertain whether differences observed between the two groups are due to the intervention or other factors that differ across the two groups. Furthermore, if the sample is too small, identification of an actual difference will not be made (statistical Type II error).

---

6 As opposed to the *external assessment* discussed in Sec. 5.
There are a number of measures that scientists can use to ensure that their control and intervention groups are sufficiently identical. In controlled laboratory experiments using animal models, e.g. laboratory mice, scientists can use animals that are bred to be identical on all relevant parameters. When this is not possible, the optimal way to ensure that the intervention and control groups are relevantly identical is to have a very big sample population and randomly assigning each individual to a group, making the study a *randomized controlled trial* (RCT, see Table 1). The sheer size of the population should here ensure that any diversity that is found in one group will also be found in the other, and the *random* allocation ensures that the researcher does not bias the groups to get a specific result. However, this method only works if the sample population is sufficiently big\(^7\), which is part of the reason why clinical trials on medicines often have such a large number of participants and part of the explanation why meta-analyses of data from several individual studies of the same research questions are generally considered more reliable than the individual studies themselves (cf. sec 4.2). If the sample population is not sufficiently big, then complete random allocation does not ensure sufficient diversity within the two groups and should thus be combined with other strategies, including statistical controls.

---

\(^7\) How big is “very big”? [that depends of the study, ask your statistician]
### Table 1: A simplified overview of central steps in an RCT

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: Identify the research question and population of interest.</td>
<td>E.g. whether gene $x$ is a causal factor for developing disease $y$ in cats</td>
</tr>
<tr>
<td>1: Identify a source population</td>
<td>The source population should be big and diverse enough to be divided into two groups, both of which reflect all relevant diversity found in the population of interest (the target population).</td>
</tr>
<tr>
<td>2: Randomize</td>
<td>Randomly allocate the members of the source population to a control- or an intervention group. Each group should reflect all the relevant diversity in the target population</td>
</tr>
<tr>
<td>3: Intervention</td>
<td>Intervene on a single variable (e.g. by knocking out gene $x$) so that it has the same value $x_i$ for each member in the intervention group.</td>
</tr>
<tr>
<td>3a: Ensure identical background conditions post randomization</td>
<td>Except for the intervention tested, all relevant stimuli experienced by members of the intervention group should also be experienced by the corresponding members of the control group to reduce the effect of confounding.</td>
</tr>
<tr>
<td>4: Collect data</td>
<td>Collect data on $y$ at least once before terminating the study.</td>
</tr>
<tr>
<td>5: Analyse data</td>
<td>Apply appropriate statistical methods to assess whether the intervention has affected a statistically significant difference in the development in the intervention group compared to the control group.</td>
</tr>
</tbody>
</table>

Assessing whether sufficient measures have been taken to avoid material errors like the ones mentioned above is one important aspect of assessing the internal validity of a source. If sufficient measures have not been taken, then this should be noted and taken into account in the final review.
4.1.2 Observational error
Even when sufficient measures against material errors have been taken, and the phenomenon of interest has been correctly generated, empirical uncertainty may still arise due to potential observational errors. Observational errors may occur in the individual measurement, e.g. because of a faulty instrument or an untrained or biased observer. But even if the individual measurements are of very high quality there may still be observational errors in the data set due to sampling errors. Sampling errors occur when the data set does not cover all of the relevant times and places related to the study. For instance, a researcher studying hen welfare in a free range system may be very good at collecting data, but if he terminates the study too soon, or only collects data on free range hens that use the outside area, thus neglecting the significant fraction that never venture out even though they have possibility, he has still made an observational error, in the form of a sample error, even though the data that he actually obtained are of very high quality.

4.1.3 Conceptual error
In addition to empirical uncertainty relating the quality and quantity of data, there may also be uncertainty in the interpretation of the data. This is part of what is often called theoretical uncertainty (Andersen et al. 2006, Ch. 10). Theoretical uncertainty can stem from both material and observational errors as they can result in data that are difficult to interpret, but it can also stem from what Allchin (2001) calls conceptual errors.

Conceptual errors include, but is not limited to, inappropriate use of statistical methods, errors in reasoning and theory entrenchment, i.e. where a researcher (or a group of researchers) are so convinced that a given theory is correct that they become blind to other plausible interpretations of their research data (cf. Fig 4). Conceptual errors may thus occur even in the fictional case where the data gathered is completely free of empirical uncertainty.

Checking for conceptual errors is often difficult as it implies critical reading of the methods section to assess whether the statistical methods applied are appropriate, critical reading of the core arguments in the source to see if the reasoning is sound and taking a step back and considering whether the interpretation of data
presented in the source is the only reasonable one. Particularly the latter step can be difficult as it requires significant imagination and understanding of theory. One of the advantages of comparing different sources on the same topic in a literature review is that it can in some cases suggest alternative interpretations of data.

In addition to the notes on the internal assessment of the arguments in the sources included in your review, it can also be useful for the later steps in the process to note where the study is placed in the *Pyramid of Evidence* (PoE). This will give you a rough indication of how strongly the arguments in the study should be weighed when compared to other studies that investigate the same research question.

### 4.2 Placing the source in the Pyramid of Evidence

The sources for your review are likely to be of different types and likely to report studies based on different methods. Most of your sources are likely to be single study reports on e.g. RCTs or some other type of study. You may also have found a systematic review article that is relevant to your research question. If the different sources point in different directions with regards to the research question you are considering, the question will rise which type of study is generally the most reliable? The PoE (Fig. 5) was developed within the medical sciences exactly as a tool for assessing the reliability of sources of different types. Making a very rough generalization, researchers in the medical sciences are primarily interested in investigating *causal relationships in a specific species* (humans). Important examples include the causes of diseases and the effects of treatments. This is reflected in the design of the PoE and should be considered when applying it outside the medical sciences (as discussed further below). That being said, the PoE is a valuable tool when facing conflicting evidence on a research question originating from the animal sciences, as these also commonly concern causal relationships in specific species albeit a specific animal species rather than humans.

![Christian Eijkman](http://www.shipseducation.net/modules/biol/beriberi.htm)

*Figure 4: Conceptual error. In the 1880's Christian Eijkman found that he could induce the disease beriberi in chicken if he restricted their diet to white rice. He also found he could cure the chickens again by feeding them small amounts of rice shells. However, Eijkman was so convinced that beriberi was caused by a bacterium that he failed to realize that his research indicated the existence of a then unknown type of nutrient: Vitamins. Later Eijkman’s data was reinterpreted and played a central role in the discovery of vitamin B₁.*

*Source: [http://www.shipseducation.net/modules/biol/beriberi.htm](http://www.shipseducation.net/modules/biol/beriberi.htm)*
The version of the PoE presented here includes eight different types of studies ordered according to how reliable each type of study typically is.

**Case reports:** At the bottom of the PoE we find reports on single events. An example could be a report on hen welfare at a specific farm. Case reports can be extremely valuable in providing a deeper qualitative understanding of complex phenomena, but since they only report on a single case, it can be difficult to draw general conclusions from them. They are thus considered the least reliable of the different types of studies used to investigate causal relations.

**Model organism and *in vitro* studies:** Just above case reports we find, in this version of the PoE, model organism and *in vitro* studies. Model organisms like mice and fruit flies are commonly used to study causal relationships either in other specific species, e.g. humans, or in a broader range of organisms, e.g. mammals. Similarly *in vitro* studies use parts of organisms placed outside their normal biological context as a model for whole organisms in their natural environment. Results obtained from both types of studies thus have to be

---

Figure 5: An example of *The Pyramid of Evidence* presenting a hierarchy of the different types of studies that can be available on cause-effect relations in a specific species. The strongest type of evidence is at the top of the pyramid.

---

8 If you google “the pyramid of evidence”, you will find that there are many different versions of it. Not all versions have the same number of “steps” and, perhaps more worryingly, not all versions agree on where in the PoE different types of studies go. Especially the ranking of animal model and *in vitro* studies differ across versions. The reason for this difference is, as will become clear below, that it is not possible to say that one type of study is better than another completely independently of the specific research question considered.
“translated” from one organism to another or from one biological context to another. This can introduce significant theoretical uncertainty, which is the main reason why these types of studies are placed so low down in the PoE.

Above model organism and in vitro studies we find first three different types of observational studies (cross-sectional, case-control and cohort studies) followed by RCTs, mentioned in Section 4.1, which is a type of intervention or experimental study. The PoE thus indicates that a well performed RCT on the species of interest is generally more reliable than an observational study when the purpose of the study is to gain knowledge about causal relations in the specific species. The reason for this is that observational studies, where the aim is to intervene as little as possible in the processes of nature to get as realistic a picture of them as possible, are prone to theoretical uncertainty due to the possibility of confounding (Andersen et al. 2006, Ch. 5). Experimental methods like RCTs have been developed specifically to investigate causal relations and gain control over confounding variables. In RCTs an intervention is (in principle) made on one specific variable on all members of the intervention group to see if the intervention causes a statistical difference in the development of another variable compared to the control group, which, as discussed in Sec. 4.1, should be sufficiently big and diverse to contain all and only the diversity found in the intervention group. This means that if the trial is done correctly, it will be easier to interpret the results, because the only difference between the control and intervention groups is that one specific variable has been manipulated in the intervention group and not in the control group. This intervention is thus very likely to be the cause of any statistical differences between the two groups. However, if the trial is not done correctly, the advantage of RCTs compared to well-performed observational studies may be lost entirely. RCTs are prone to material errors partly because they require big sample populations that can be difficult (and expensive!) to manage and partly because it can be difficult in practice to make sure that interventions are made on only a single variable during the course of the study. It is thus important to note that although a good RCT can often be more reliable than an observational study, this advantage is easily lost due to the vulnerability of RTCs to material error.

Cross-sectional studies: Of the three types of observational studies included in the version of the PoE presented here, the cross-sectional study is the simplest. Here the aim is to take a “snapshot” of the distribution of a certain set of variables in a given population at a given point in time. These data can for instance be used to find correlations among the variables measured, e.g. whether barn hens have a higher risk of feet problems than hens in other housing systems. However, the many challenges related to distinguishing between correlation and causation (Andersen et al. 2006, Ch. 5) means that these data are not very reliable when it comes to investigating causal relations.

---

9 The difference between experimental and observational studies is discussed further in Ch. 6 of Andersen et al. (2006), which also contains a more detailed discussion of RCTs.
**Case-Control studies:** In case-control studies we seek to uncover one or more causal factors for a given effect by comparing one group of subjects in which the effect is present (the case group) to another group in which the effect is not present (the control group). The aim is to identify interesting differences between the two groups that can explain why the effect is present in one group and not the other. This gives case-control studies some advantages compared to the cross-sectional studies when it comes to identifying causal relations, but they also have weaknesses. First of all the study design prevents us from saying anything about how likely it is that a causal factor will produce the effect we are studying. At best we can learn that the identified factor *can* produce the given effect.

Furthermore case-control studies are particularly vulnerable to a kind of material error known as spectrum error or *spectrum bias*. Because we include subjects in a case-control study based on whether the effect is present or not, there can be tendency in these studies to include only the clear cut cases. For instance, in a case-control study focusing on the cause of a certain disease in layer hens, there can be a risk that only hens with very clear symptoms of the disease and hens that are very clearly healthy are included, because these are the easiest to place in the case and control groups respectively. However, this means that we could end up comparing only the two extremes of the spectrum between being very healthy and suffering very badly from the disease, which could mean that the results we get are not easily translatable to the less severe cases.

**Cohort studies:** Contrary to both cross-sectional and case-control studies, cohort studies always include a time component, looking either backward (called a *retrospective* cohort study) or forward (called a *prospective* cohort study). The subjects participating in a cohort study are chosen differently than the participants for a case-control study. In a cohort study we select the participants among the subjects in which the effect we are interested in is not yet present. From this group subjects are included in the study based on the presence of one or more factors that we suspect are causal factors for the effect. The subjects are then followed over time to determine if the effect is more likely to occur in subjects in one group, than in subjects in another group. Because of the time component, and because we make sure that the effect was not present in any of the subjects from the beginning, we know that the suspected causal factor was present *before* the effect. This makes it easier to distinguish between cause and effect in cohort studies compared to cross-sectional studies and case-control studies. Furthermore cohort studies can tell us more about how likely the effects is to occur given the presence of the causal factors that we are investigating.

The study by Sherwin and colleagues (2010) on the welfare effects of different layer hen housing systems discussed in Section 2.1 is an example of a very elaborate prospective cohort study comparing four different cohorts; one for each housing system. The study began, when the hens in each cohort began laying (age 16 to

---

10 The term bias is used in slightly different ways in different fields. Here we follow the broad definition in Andersen et al. (2006), Ch. 6, where the term is synonymous with the “systematic error”, i.e. an error that always affects the result in the same way.
20 weeks) and followed them for up to 50 weeks, making regular cross sectional measurements of various welfare indicators. The data from each cohort were then compared to yield the result that the overall welfare in furnished cage systems was generally higher than in other housing systems.

Although cohort studies have advantages over other types of observational studies they are still placed lower in the PoE than experimental studies like RCTs because of the increased control of confounding variables that a well performed experiment can offer.

Reviews: At the top of the PoE we find thorough review articles (preferably systematic reviews with meta-analyses). Review articles are considered highly reliable partly because they rely on the largest and most diverse data sets and are thus less likely to include e.g. material and sampling errors than the individual studies that are part of the review. Furthermore they are less likely to suffer from theory entrenchment as they can draw on the theoretical perspectives of many different studies. Researchers and regulators that need to make important researched based decisions, e.g. on whether a new drug is safe and effective enough to be released unto the market, therefore tend to rely heavily on systematic reviews and not to the same extent emphasize individual studies, even though they may indicate a different more worrying result. However, reviews are not necessarily more reliable than single studies. The discussion in Sec. 2.1 illustrates that a single very well performed study placed high up in the PoE can weigh at least as heavily as a qualitative review of many studies that are either less well performed or placed lower down in the PoE.

4.2.1 Limitations of the Pyramid of Evidence
The PoE can be used to help you assess how much emphasis you should place on an individual study in cases on conflicting evidence. If you have two sources that contradict each other on the answer to a given research question related to a causal relation in a given species, you should, generally, rely on the one that is placed highest up in the pyramid. The “generally” is extremely important here. The ranking of the different types of sources in the PoE is based on a number of assumptions including:

1. that the studies are all very well performed,
2. that they deal with exactly the same research question, and
3. that the research question is about a causal relation in a specific species.

If one of these assumptions is not fulfilled, it can mean that the ranking presented in the PoE will not be valid. This is why the PoE can never be used on its own, but must always be accompanied by a more general

---

11 This can sometimes be difficult to explain to non-experts. Many news stories start with the observation that “researchers have found” that something we use in our daily life might be toxic/carcinogenic/etc. Commonly this claim is based on single study, and the challenge for researchers and authorities is to explain that the reason why there is no immediate reaction to the new research is that decisions to e.g. ban a chemical or product are rarely based on single studies, but on systematic reviews on the safety of said chemical or product and that the new study is not sufficient to tip the balance in favor of banning.
internal assessment as sketched in Sec. 4.1. Otherwise you cannot be sure that assumption 1 is valid. Studies that do not deal with the same research question (and thus do not fulfill assumption 2) cannot be compared solely based on their internal qualities, but must also be judged on their relevance for the specific question that you are interested in. Here it might turn out that a cohort study on exactly the question you are interested in should be weighed higher than an RCT that is slightly off topic (cf. Sec. 5.2). Finally, although very many research projects within the biological sciences concern causal relations in one way or another, this is not always the case. It is certainly not the case that all projects concern a specific species. As we saw, the ordering of the different observational studies and RCTs was based heavily on considerations on the ability of these types of studies to reliably reveal causal relations. Cross-sectional studies were thus placed low in the PoE because it can be very hard to distinguish causation from mere correlation based on cross-sectional data. However, cross-sectional studies can be very useful for other purposes. If your research question concerns e.g. the distribution of a certain trait at a certain point in time, then a cross-sectional study will be perfectly suited for the job.

If your study does concern a causal relation, but in a group of species, e.g. rodents or fish, rather than a specific species then the categories presented in the PoE become harder to apply. Particularly the “model organism” category becomes problematic. Questions about mammals can thus be investigated through RCTs on mice. But how should such studies be categorized? Should the mice be interpreted as a) a model organism, placing the study low down in the PoE? Or should such studies rather b) be categorized as RCTs with a highly unrepresentative sample population, implying that assumption 1 is not fulfilled and thus that the PoE does not really apply? Option a) will be appropriate in many cases, but case to case judgments may have to be made.

Once you have placed each of your core sources in the PoE and made thorough notes on the internal assessment of each core source, you are already a long way towards getting an overview of the sources included in your review. The next step is to consider the possibility of publication bias.

### 4.3 Publication bias

So far we have discussed how to assess the sources that you have been able to find in your literature search. However it is also important to consider, whether there are studies that have been performed, but that you have not been able to find, because they have not been published.

In an ideal world, the results from all well performed studies are reported to the scientific community so that they can be used in the further development of our common knowledge. This goes not only for the studies that yield clear, surprising results, but also the studies that yield less surprising results or no clear results at all. If studies are performed and not published they do not benefit the broader scientific community. Furthermore, if there is a systematic tendency to publish only research that yields a certain type of result, e.g.
results that confirm the hypothesis tested, it will introduce a bias in literature reviews based on the published literature. What we might call publication bias.

Unfortunately we do not live in an ideal world, and publication bias is a very real phenomenon. Studies that confirm the hypothesis of the authors are more likely to be published than studies that disconfirm the hypothesis of the authors, which are in turn more likely to be published than studies that do not yield a (statistically) significant result and thus neither confirm nor disconfirm the hypothesis of the authors (Stern & Simes 1997). Why this tendency appears is not entirely clear, but one important factor is that researchers are less likely to submit manuscripts reporting negative and non-significant results to academic journals, because such manuscripts are less likely to be accepted by the editors of said journals, partly because the readers are less interested in reading and citing such results.

The consequences of publication bias for the reliability of literature reviews based solely on published studies are significant. If studies confirming e.g. the therapeutic effect of a given drug are more likely to be published than studies of similar quality that dis-confirm or say nothing conclusive about the therapeutic effects of the same drug, then a review of published literature on the therapeutic effects of said drug is likely to be biased towards confirming that the drug actually works, even though the sum of all performed studies may not actually show this.

If the evidence presented in your sources points strongly towards confirming a given hypothesis, it is thus worth considering whether this tendency could be due to publication bias, rather than an actual tendency in the total evidence gathered. Circumstances that should raise suspicion include:

a. *If the negation of the hypothesis is not an interesting scientific result.* In practice, scientific results often need to have some ‘news value’ in order to be published in high ranked journals. A study that set out to confirm a highly controversial hypothesis, but ended up confirming what everyone thinks is thus less likely to be published than studies of similar quality, but with a higher news value.

b. *If there are strong (economic, social, etc.) interests in confirming the hypothesis.* Classic examples are drugs or other kinds of treatments that medical companies have invested a lot of money in developing. This can in some cases lead researchers who are economically dependent on companies to deliberately withhold negative results from publication, but it can also lead researchers to unintentionally make conceptual errors and misinterpret high quality data indicating a negative result as flawed data that are unfit for publication.

---

12 From a philosophy of science perspective this is an interesting result. Karl Popper famously argued that science was not about verification (showing that you hypothesis is correct). Rather the scientific method consists in purposely trying to disprove (falsify) you hypothesis ((Popper 1968) see also Andersen et al. (2006), Ch. 6). It seems that this idea is not reflected in the actual publication practice of science.
It is often very difficult to test suspicions of publication bias. In some cases, including RCTs on medicines, there are requirements that require that all planned projects should be registered prior to the data collection and that the raw data must be made available for meta-analysis regardless of whether the studies are eventually published or not. But even if it is in principle possible to search for unpublished data, it is not necessarily possible in practice within the timeframe of e.g. a bachelor project. If you suspect publication bias you should therefore start by discussing this possibility with your supervisor. If your suspicion remains, and remains untested, it is important that you make your suspicion explicit in your review, and take it into account when writing the review.

Once you have completed the internal assessment of your sources, placed them in the PoE, and considered the possibility of publication bias, you are finally ready to begin comparing the sources to each other.

Section 4 Reflection questions
1. What could be other examples of material and observational error within your field?
2. Can spectrum bias occur in cohort studies?
3. Consider a paper you read recently. Where would you place it the PoE?
4. RCTs on model organisms are placed much lower in the PoE than RCTs on the target species. Why?
5. Discuss advantages and disadvantages of excluding e.g. everything below cohort studies from your review in order to be sure that your review relies only on the most trustworthy sources.

Synthesizing your sources
As mentioned in Sec. 2 it is important that a review provides a) an overview of the sources by grouping them e.g. according to whether they support or contradict a specific hypothesis, b) discusses the reliability of the individual sources, and c) the conclusions that can be drawn by comparing the sources to each other. Furthermore the good review may also d) point to areas of significant scientific uncertainty that are in need of further research.

A common first step in preparing the review is therefore to group the sources into relevant categories. Often it will be relevant to group the sources according to whether they support the same conclusion on your research question, or one of its sub-questions.

Your sources will generally not all consider exactly the same research questions, and the questions they consider will in many cases be different from the specific research questions considered in your review. This means that you for each source must assess how it is relevant to your specific research questions. This is known as the external assessment of the source. This also means that you cannot simply conclude that the sources that have the strongest internal assessment also provide the strongest evidence relative to your research question.
Lay and colleagues faced these challenges when they compiled their review on hen welfare in different housing systems (discussed in Sec. 2.1). Very few of their sources addressed the general question about the relations between hen welfare and housing system. Rather, they focused on more specific relations e.g. the relation between housing system and stress, or housing system and foot health. To get an answer to their research question Lay and colleagues thus had to combine the results from the different sources into an answer to their own research question.

Often a significant task for the reviewer is thus to break the original research question down to different sub-questions, and show how the different sources can help answer these sub-questions. When trying to synthesize your sources you may for instance find that some of your sources address some of your sub-questions directly, but have significant scientific uncertainties related to them, leading you to pose new questions. If you are lucky you find other studies that address these uncertainties and adding these to your review may thus enable you to argue more strongly for the conclusions reached in the more directly relevant studies. If you are not able to find answers to all of your most relevant sub-questions then this is also a result: you have found a gap in the literature in need of further research.

If you group your sources according to whether or not they support a given hypothesis, it will be important to go into an analysis of why disagreements occur across the two groups. Commonly, disagreements are due to differences in methodology, slight differences in the framing of the research questions or disagreements on the specific meanings of key terms. As illustrated in the following example.

5.1 Grouping your sources: an example
A recent review article by Protopopova (2016) covers just over a hundred studies on the effects of sheltering on the welfare of dogs. Despite there being quite a few studies on the topic, many of which indicate that dogs in shelters generally have lower welfare than dogs in private homes, the author concludes that no clear conclusion can be drawn from the sources available due to the substantial scientific uncertainty that surrounds the topic.
As we know, animal welfare can be defined and studied in many different ways and Protopopova has therefore chosen to group her sources first according to methodology and then according to whether or not they indicate that sheltering has a negative effect on the welfare of the dogs. One group of studies that use very similar methods assumes that increased cortisol-levels in the dog can be used as an indicator of stress and thus of reduced welfare. These studies then proceed to measure the effects of sheltering on the cortisol-levels of dogs. Another group of studies relies on the association between stress and reduced immune function and proceeds to investigate whether dogs in shelters have reduced immune function compared to dogs in private homes. A third group of studies focuses on the behavior of the dogs and look for e.g. repetitive behavior, fear behavior and abnormal activity levels as indicators of reduced welfare. Within each group there is conflicting evidence on whether or not sheltering has a negative effect on the welfare of the dog beyond the first few days. A major challenge within the first two groups, according to Propotopova, is the substantial theoretical uncertainty related to the interpretation of data. For instance, the data on cortisol levels are of fairly high quality and indicate that dogs generally experience an increase in cortisol level when they arrive at a shelter but the cortisol level then tend to gradually decrease to a level similar to those of dogs in private homes after a couple of weeks. The question is how this should be interpreted. Is the reduced cortisol level an indication of reduced stress over time, or rather an indication of an overload and subsequent dys-function of the stress hormone regulation system in the dog, which would be an indication of chronic stress and reduced welfare? It is, according to Protopopova (2016, p. 97), not possible to distinguish between these two possible interpretations based on the currently available data.

The behavioral studies also have theoretical uncertainties related to them. In addition, Protopopova points to a significant source of empirical uncertainty in the behavioral studies. Many behavioral studies are based on observations of repetitive or stereotypical behavior. When comparing studies on the prevalence of stereotypical behavior in dog shelters, it is of course important to be sure, that the authors of the different studies agree on when stereotypical behavior has been observed, e.g. how long a dog must have been observed pacing along the fence of the kennel for it to qualify as stereotypical behavior. Unfortunately this is not always the case. Protopopova points out that the different studies operate with somewhat different operational definitions of when stereotypical behavior has been observed. This makes it more difficult to compare the studies, and interpret the differences in results, as it becomes unclear whether the differences are

Figure 6: Pit Bull in a shelter. Dogs in shelters will sometimes develop stereotypical behavior, where they e.g. pace along the fence of the kennel or chase their own tail. However, observers do not always agree on when stereotypical behavior has been observed. How long should a dog e.g. pace along the fence for its kennel to be categorized as displaying stereotypical behavior? Source: http://maxpixel.freepikpicture.com/
due to interesting differences in the kennel environments or less interesting differences in operational definition of stereotypical behavior.

Because of the substantial uncertainties surrounding the existing research on the effects of shelters on dog welfare, Protopopova concludes her review by pointing to areas that most urgently require more research and attention.

Protopopova’s review is a fairly standard example of a non-systematic review article published in an academic journal. It provides an overview of the field by grouping the sources in a relevant ways, points out conflicting evidence, and discusses the sources of the disagreements that can be found in the literature. In this way, it provides other researchers with suggestions for interesting research questions and, importantly for the author, a chance for the author to show that her own research contributes importantly to a relevant debate.

Protopopova concluded that it was not possible to draw a reliable conclusion based on the evidence she had reviewed. In Sec. 2.1, on the other hand, we found that it was possible to conclude on the welfare effects of different housing types for laying hens based on the available evidence even though there was also conflicting evidence in this case. What made the difference?

5.2 Considering uncertainties

Protopopova was unable to draw a clear conclusion both because of the amount of scientific uncertainty surrounding her research question, but also because of the way the uncertainty is distributed. There are substantial theoretical and empirical uncertainties surrounding both the studies indicating that sheltering has a negative on the welfare of dogs and those that indicate that there is no effect. This makes it very difficult to draw a clear conclusion.

In the case of hen welfare discussed in Sec. 2.1, there was also scientific uncertainty surrounding the various studies considered in the review. However, because the best studies pointed in the same general direction – that furnished cages provide some of the welfare benefits of non-cage systems, while at the same time maintaining the welfare benefits of the cage – and because the uncertainties related to the individual studies pointing in this direction could partly be resolved by considering the sources in combination and thereby reducing the uncertainty even further, it was in this case possible to draw a reliable conclusion.
More generally, it is thus possible to draw reliable conclusions from conflicting evidence if:

1. It can be argued that there is significant uncertainty surrounding the evidence speaking *against* the conclusion, AND
2. It can be argued that the uncertainty surrounding the combination of evidence speaking *for* the conclusion is sufficiently small to allow for a conclusion to be drawn.

The development of these arguments will to a large extent be based on the internal assessment of the individual sources as sketched in Sec. 4. However, the internal assessment must be supplemented with the external assessment focusing on the relevance of the specific research questions asked in the sources relative to the specific research question considered in your review and the research questions considered in the other sources used.

In some cases your sources will have to reinterpreted to be used as partial answers to one of your sub-questions. In discussions of the effect of sheltering on dogs and other pets it is for instance not uncommon to see references to results from studies on zoo animals. This can be very useful, but one must also consider the uncertainties that are introduced by such reinterpretations. Dugongs are not dogs, and reinterpretation from one species to another and from one environment to another always introduces uncertainties. In Sec. 4.2 we mentioned this in relation to studies performed on model organisms and *in vitro*. We saw that such studies are placed fairly low in the PoE exactly because of the theoretical uncertainties that arise in the translation across species and/or environments. The same uncertainties of course arise if the translation is performed by others than those who performed the study. This means that when e.g. an RCT or a case-control study performed in a zoo is reinterpreted in a review focusing on pets, then these studies should be considered as model organism studies and weighed accordingly (cf. Fig. 5).

**Section 5 reflection questions**

1. What are the advantages of grouping sources first according to methodology and only secondly according whether or not they support your hypothesis as Protopopova did? What are the advantages of doing it the other way around?
6 Summary: The essential steps in making a literature review

We have now discussed some core steps in making a qualitative review of an extensive but incomplete set of sources aiming to find out what the sources can tell us about a specific research question. This is typically the kind of review that a student will be asked to perform when writing the bachelor project. The process is summarized in Fig. 7 (see next page). You can use this figure as a process guide when writing your own literature reviews. It will also help you go back to the previous sections in this note for input on important things to remember when going through each step in the process. In the actual process of writing your own review, you may of course find that some of the issues discussed above are not relevant to your specific review, or that your will have to go through the process in a slightly different way than sketched in Fig. 7, e.g. go back and perform additional literature searches during the synthesis. All of this is perfectly fine, as long as the end result is a review that fulfills its intended purpose to the widest extent possible. As previously mentioned this at least implies that the review should go beyond being a simple listing of previously obtained results. The review should add something new, e.g. by showing how the previously obtained results can be combined to gain new or more thoroughly justified knowledge.
**Outline of a potential review process**

- **Research question**
- **Initial literature search**
- **Overall purpose of the study**

**What is the purpose of your review?**

**Do you need to make a more systematic literature search?**

**Do you want to include all your sources in your review?**

**Define exclusion/inclusion criteria**

---

**Internal assessment**

- Are there significant uncertainties in your sources?
  - Randomization
  - Sampling error?
  - Spectrum bias?
  - Theoretical uncertainty?
  - Etc.

- Place your sources in the PoE

- Could there be publication bias? If so, how should this be handled?

---

**Synthesis**

**Qualitative review**

- Group your sources
- External assessment
- Discuss conflicting evidence
- Conclude

**Quantitative review**

- Extract data
- Perform meta-analysis

---

**Figure 7:** Outline of a potential review process. Based on a research question, the overall purpose of your study (publication, BA-project, etc.) and an initial literature search, the purpose of the review is identified. A narrative review is a possibility, but is not discussed here. For a more thorough review, there may be a need for a supplementary literature search before proceeding to considerations about inclusion/exclusion. Once a set of core sources has been identified the internal assessment of each source proceeds as sketched in Sec. 4, followed by synthesis as sketched in Sec. 5 for the qualitative review. The details of the quantitative review are not discussed here.
7 Perspectives

This note began with a quote from Isaac Newton, perhaps the most celebrated scientist in history: “If I have seen further it is by standing on the shoulders of giants”. The quote reminds us that even the greatest scientists rely heavily on the existing literature when doing their research. Figuring out what the existing scientific literature can tell us about a given research question is an important step in any scientific project, but not necessarily an easy step. Particularly if there is conflicting evidence. But by carefully considering the internal and external validity of the available sources it will in some cases be possible to navigate through the conflicting evidence and reach conclusions that can be more reliable than the conclusions presented in the individual studies. Literature reviews thus not only serve as an important first step in the research process; they serve important roles later on as well.

One of the central themes in discussions on methodology in philosophy of science courses is the possibility crucial experiments i.e. single experiments that confirm or disconfirm a given hypothesis. Popular accounts of science contain plenty of stories of crucial experiments that convinced the world that a given hypothesis was correct. Think of the story of Galileo dropping spheres of different masses from the Leaning Tower of Pisa to show that they hit the ground simultaneously, thus disproving central parts of Aristotle’s physics, the story of Pasteur’s experiments that allegedly settled the debate over the possibility of spontaneous generation (Fig. 7), or the story of Mendel’s pea experiments proving what later became known as Mendel’s laws. In reality however, single experiments are rarely sufficient for new spectacular scientific results to be generally accepted in the scientific community, and there are very good methodological reasons why this is exactly as it should be. The many possibilities for error in the process of testing a hypothesis means that we should at least require that results of single tests are reproduced by others (perhaps through different methods) before accepting a new daring hypothesis or rejecting a hitherto well established hypothesis (Andersen et al. 2006, Ch. 6). Several different studies are thus generally required to reach such a conclusion, and the results from these studies need to be weighed against each other in order to build a stronger, more nuanced justification for the conclusion. This is exactly the purpose of the hypothesis testing literature review. The different types of literature reviews thus not only serve as a first step in a research process. They also serve as one of the last.
Section 7 reflection questions
1. Discuss the functions of literature reviews can have in reaching reliable scientific results.
2. Discuss the functions of literature reviews in can have in policy making.

References


