Detecting plagiarism using Similarity Check

Similarity Check "helps editors compare the text of submitted papers for similarity" (https://www.crossref.org/services/similarity-check/).

When a document is checked in Similarity Check, it is compared with the content of this database which is made up of published and unpublished documents, including over 40 million research articles, conference proceedings, and e-books from scientific, technical, and medical publishing. It includes material behind journal paywalls that would not be available from a simple internet search.

Getting started with Crossref Similarity Check

Cochrane Review Groups (CRGs) are encouraged to use Similarity Check via the licence held by the publisher of the Cochrane Library, John Wiley & Sons. Wiley provides each CRG with access to Similarity Check (free of charge). For a user name and password, Managing Editors can contact Tony Aburrow (taburrow@wiley.com), Cochrane Editor, Wiley.

What and when to check

CRGs are encouraged to, at minimum, check at least a portion of text for all protocols and reviews (including updates) when initially submitted to the CRG.

There are different stages in the editorial process where Similarity Check screening could occur (see *Table 1*). CRGs may wish to screen more than once, or they may wish to screen at a particular time, such as before peer review, or where the writing styles varies within a single document.

Table 1. Different stages in the editorial process where Similarity Check screening could occur

Stage	Document	Recommended sections to screen[a]
Title	All Review Proposal Forms	All text excluding references
Protocol	**Initial submission of protocol**	Background[b], Methods[b]
	All resubmissions of revised protocols	As above
	Substantively updated protocols (i.e. new citation version)	As above
	Final version for publication	Screening not recommended at this stage
Review	**Initial submission of review**	Abstract, Plain language summary, Background[b], Methods[b], Results, Discussion, Authors' conclusions Omit (1) matches to the published protocol from the similarity report and (2) references
	All resubmissions of revised reviews; or review 'amendments'	Where changes have been made to the text
	Updates (initial version and revisions)	Abstract, Plain language summary, Background[b], Methods[b], Results, Discussion, Authors' conclusions Omit (1) matches to the published protocol; (2) published previous versions of the review from the similarity report[c]; and (3) references
	Final version for publication	Screening not recommended at this stage

[a] While it is possible to check an entire document for similar text, sections of a Cochrane Review, such as the methods, characteristics of studies tables, and references sections, are likely to give a high similarity score due to the nature of their content.
[b] Some Cochrane Review Groups may recommend the use of template text for the Background or Methods section. If so, the authors should have made a note of this within the protocol or review. See 'Special circumstances for Cochrane Systematic Reviews' for more information.
[c] It is possible to do this in Similarity Check; see *Table 3*.

How to check

Similarity Check provides a similarity score, which indicates the total amount of text that matches text in other sources. There are two steps to using Similarity Check: (1) an automated step in which Similarity Check runs the online comparison; and (2) a manual step for someone in the Cochrane Review Group to interpret the report results and decide on next steps; see *Table 2*. These two steps combined can take from 5 minutes to 2 hours, but it is usually around 15 minutes. Similarity Check provides a list of resources for using the software: www.ithenticate.com/resources/customer-training/.

Table 2. Overview of Similarity Check process

Automatic process	Similarity Check finds and highlights overlapping text between manuscript and published material
	A similarity score is generated
Manual process	Similarity Check report reviewed
	Determine severity of plagiarism
	Decide on action to be taken

Once logged into Similarity Check, there is the option to submit different file types for screening. It is not recommended to submit the full version of the document because it may be very long and will include sections that have little value in being screened (e.g. references); see *Table 1*. Therefore it may be easier to select specific sections of the protocol or review to be screened. There are three possible approaches:

1. Prepare a new document by cutting and pasting specific sections of text into a new document and save as one of the following file types: plain text, MS Word, PDF, RTF, PostScript, HTML, or XML.
2. Use the Cut and Paste upload option in Similarity Check.
3. From RevMan, using software installed to print to PDF, select the required sections and print and save as a non-RevMan PDF file type. Similarity Check does not accept RevMan file types (i.e. *.rm).

By default, Similarity Check includes the optional settings to exclude quotes (i.e. text within quotation marks), reference lists, and/or “small matches” of text to avoid false positives in the similarity index. However, while it is possible to request references to be excluded from comparison using Similarity Check, this does not always happen and it is preferable to upload a file without this section. It is not always advisable to exclude “small matches” to text because small matches could be direct quotes that need quotation marks and citations.

Similarity Check has an option to include a simultaneous Internet search (called “websearch” in Similarity Check) in addition to the standard iThenticate database search. This extends the Similarity Check comparison to include content not included in the iThenticate database, such as Wikipedia, and presents the collated results. This option should be used routinely.

When matches are identified in a report, Similarity Check has an option to exclude one or more matching sources. As described in Special circumstances for Cochrane Systematic Reviews, a high percentage of overlap would be expected between a protocol and review, and a review and an update. This functionality allows the user to exclude the protocol or original review, for example. This functionality may become less useful as the number of times a Cochrane Review is updated as the number of exclusions that need to be made increases. See *Table 3* for the types of Similarity Check reports where this functionality is available.

CRGs should agree which editorial staff member(s) should be responsible for running the Similarity Check reports, interpreting the results, and deciding on next steps. For example, an Assistant Managing Editor/Managing Editor, Information Specialist or administrative assistant could run a document through Similarity Check and generate a report. The results of the report should be considered by the CRG’s Managing Editor and/or Co-ordinating Editor and any action to be taken decided upon.

Similarity Check reports

There are different modes of reporting in Similarity Check (see *Table 3*) some of which display different information. The Document Viewer is the default setting and shows the best matches for text in a submitted document (see *Figure*). The Document Viewer report has two pieces of information that will guide the editorial team to have no cause for concern or to decide if any action is needed:

- It highlights any overlapping text and shows you where it comes from, and how many words are overlapping in each instance (number of words is more informative than the percentage overlap, which is also provided). The editorial team can review all instances within the document.
- The Document Viewer report will include a similarity index score. Similarity Check’s similarity index should not be used as an absolute measure of whether significant overlap exists, but rather as a signal to have a closer look at the text. The score is a percentage of text that it has identified as an overlap with one or more other sources. A low score means less overlap and a high score means more overlap.

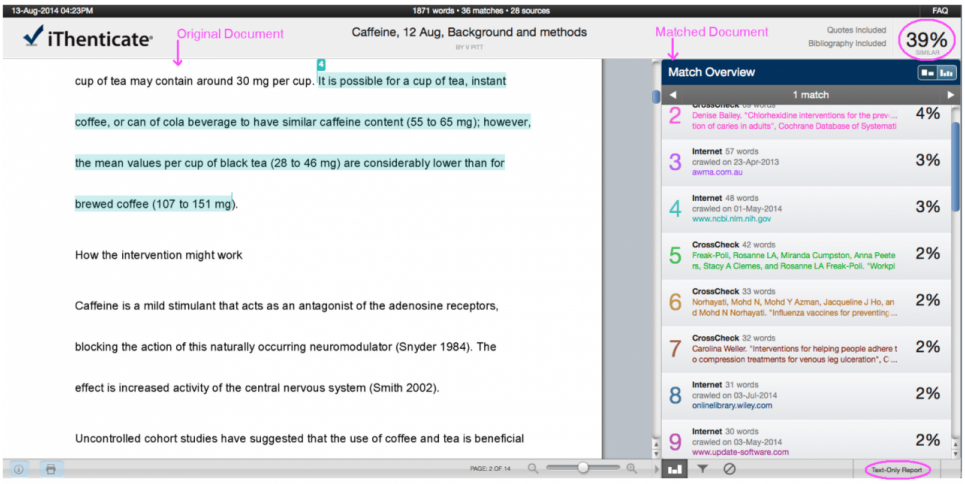
For further information on using Similarity Check, please see the official Similarity Check user manual (via TurnItin). For further information about the similarity score, see the iThenticate website.

Table 3. Types of Similarity Check reports (www.ithenticate.com/training/dv-walkthrough)

Document Viewer	Default report; a detailed report that uses colour coding to compare texts, and hyperlinks to allows user to review matches. You can exclude particular sources in this mode.
Similarity report	Displays matching sources side-by-side with sampled text. You can exclude particular sources in this mode.
Content tracking	Enables users to see if matches were manually excluded, or if there are more than one match for the sample, and ranking of proportional match in the report. You can exclude particular sources in this mode.

Summary report	Same information as the similarity report, but it displays matching sources above the document
Largest matches	Ranks sample according to the word count and percentage of words that match a string of words.

Figure. Example Similarity Check Document Viewer



This Similarity Check Document Viewer report shows the document being checked on the left side, highlighting matching text (in this example in red, blue and green), and the context of the matching text in the match document (Spirit MJ et al) on the right side. In this example the highlighted text in red and green match other sources than the text in blue and are not shown.

The “Document Viewer” is the chosen reporting mode. Clicking on the “Text-Only Report” button will change the display to other reporting modes, which are detailed in *Table 3*.

The “Similarity Index” applies to the entire document being checked and indicates the percentage of text from the entire document which overlaps with identifies sources (matched documents) and is shown in the upper right hand side of the report.

Figures and images

Editorial teams should be aware that Similarity Check will not identify any plagiarized figures or images, such as line drawings and photographs. Also see Copyrighted images: using in articles published in the CDSR.