Appendix B: Quality assessment tool

Non-trial quality assessment

Overall assessment of quality in and between studies:

Use the below table to make a qualitative assessment of quality in both individual papers, and across the literature being reviewed. Use the details of each domain below to aid assignment of quality.

Quality	Interpretation	Within a study	Across studies
High quality.	No detected issues, or issues unlikely to seriously alter the results.	High quality for all key domains.	Most information is from studies with high quality.
Medium quality.	Issues that raise some doubt about the internal validity of the study.	Medium quality for one or more key domains.	Most information is from studies at High or medium quality.
Low quality.	Issues detected seriously weaken confidence in the internal validity of results.	Low quality for one or more key domains.	The proportion of information from studies at low quality is sufficient to affect the interpretation of results.

Reference: Adapted from Cochrane handbook for systematic reviews table 8.7a. See: www.handbook.cochrane.org

Domain 1: Quality of the data

This domain is concerned only with the quality of data used in each article, and how this can affect internal validity. Authors' approach to analysing the data is dealt with in the methodology domain. The following issues should be considered, and a judgement on the quality associated with data should be based primarily on them. Issues with the data not covered below should be described in detail in order to incorporate them into the review.

For secondary data:

- Does the study acknowledge and address missing data issues?
- Does the study address the representativeness of the data used?
- Does the study address issues or problems with the consistency of data collection, trustworthiness of the data or any other issues that could bias the secondary data?
    - An example: The authors control for armed conflict to account for the non-trade related effect on mortality rates. In that case, do they consider quality or even existence of data during periods of conflict?
- If some data used in the study was composed or calculated by authors for use in the study: Was this data calculated or composed in a reasonable way that is unlikely to affect research results?
    - An example: In a study using spatial statistics and country level data, are the calculated distances between countries based on the centre of each country or distance of capital cities? Is the choice between these important? Are the authors' choices on this likely to affect the results?

For primary data:

- If the data is on an individual level, was the sampling method random?
  - If the sampling was not random, is the sampling method likely to affect the results?

- If the data is on a group level (e.g. by workplace, family group, geographical location and so on), do the authors consider that behaviour within a cluster can be correlated with that cluster (i.e. clustering effects)

- Do the authors discuss the problems that their data collection method could create? Are the effects of unique events and measurement error discussed?

Domain 2: Quality of data approach and analysis method

This domain is concerned with threats to internal validity brought about by the approach to the data the authors take, the selection of applicable analysis methodology, and the implementation of that method. Due to the topic area of this review, methodology and data are not consistent throughout selected studies. Therefore, for each statistical analysis method that could reasonably be applied to the available data, information on typical methods of showing internal validity is provided below. Reviewers should select the headers that are relevant to the article under review. With consideration to this information, reviewers should make a judgement of data approach and analysis method quality. We do not view the use of multiple methodologies as an indication of internal validity unless they are used for sensitivity analysis or some form of robustness testing. Multiple methods should be treated as such by referring to the multiple relevant headers if they are not used to test robustness of primary findings.

Knowledge of the method(s) in question is necessary to inform judgement in this domain. With some simpler methods such as Ordinary Least Squares (OLS), the Gauss-Markov assumptions are likely to be violated when looking at country level trade and population statistics. Omitted variable bias, for example, is likely to be present in country level regressions on these data due to the multitude of potentially relevant variables. Also, endogeneity, multicolinearity and other issues are likely. Therefore, it would be difficult to confidently state that quality is high in an OLS based study of the relation between trade and health unless a large range of control variables and some means to control for other issues was used. With that example in mind, please judge the quality due to data approach and methodological implementation to be high, medium or low using the following information on typical approaches that account for bias.

Approach to data

- Do the authors consider and address unobservable differences between clusters (countries, regions, individuals and so on)?
  - Example: Using dummy variables for clusters or selection of methods which take this issue into account (e.g. fixed or random effects models)

- If the authors consider unobservable heterogeneity to not be an issue, is this justified?
  - Example: By referencing other works which show the issue to be unimportant

- Do the authors control for confounders, not just variables considered in the hypothesis?
  - Example: use of a correlation table to indicate the relationships between variables that in all likelihood have a complex relationship with multiple factors.

Methodology

Below is a list of typical methods to internally validate analysis methodologies. Please consider these steps in judgement of quality.

Regression discontinuity designs:

- Are the participants blinded **or** not able to amend the control?
- Do the authors show that the difference in characteristics between the control and intervention groups is small?
- If the design is such that a point in time is the beginning of the treatment, (e.g. before and after a country joins a free trade agreement), is there data shortly before and after the treatment?
    - Example: There may be a quality if data from 5 years before and 5 years after the exposure is used, since the effect of the treatment could be very short term, and an exogenous factor entered the situation in the meantime.
- Do authors control for exogenous factors affecting the outcome through regression?

Instrumental models

- Are the instruments F significant (i.e. is F≥10 in the instrumental model)?
- If the instrumental model does not use the Heckman procedure – are the individual identifying instruments significant?
- If the research uses the Heckman procedure, are the identifiers reported and significant?
- Are at least 2 instruments used, and do authors report the results of over-identifying tests? (not always essential to score high quality)
- Do the authors conduct a qualitative assessment of the exogeneity of instruments? Is the instrument exogenously generated?
- Do the authors control for confounders, and are these controls likely to be affected by participation?

Ordinary Least Squares (including reformulations that do not fit into other headers) and MLE estimations (e.g. logit, probit)

- Are the Gauss-Markov assumptions satisfied? If not, is the violation likely to bias the results? Are issues with the assumptions addressed?
- Do authors control for confounders in their regressions?
- Do the authors use proxies to account for unobserved heterogeneity
    - Example: using dummy variable for country in a multinational analysis (i.e. as a substitute for using a fixed effects model)
- If the design is quasi-experimental, does participation affect the control group?
- If the design is quasi-experimental, are the distributions of covariates shown to be balanced across groups?

Fixed effects, random effects and difference in difference

- Do authors control for time variant characteristics?
- Do the authors test the robustness of their model? Is their method for doing this applicable to the data they use?
- Was a Hausman test used to test the relative internal validity of Fixed Vs Random effects models?
- If the design is quasi-experimental, are there low levels of attrition (<10% is acceptable for this review)

Matching estimators

- If the primary method is propensity score matching, is the 'caliper width', or propensity score matching range mentioned? Do you consider it narrow enough? (0.1 is usual)
- Are any individuals from one group are matched with large numbers from the other?

Hypothesis testing – (e.g. t-tests and non-parametric testing)

Note: If bivariate testing is used, please consider your answer for quality in the data. If testing is bivariate, the data used must be adjusted for confounders in some way.

- Is the distribution of covariates is demonstrated by authors?
- If t-tests are students', is evidence provided showing data is normally distributed provided? Otherwise, are non-parametric t-tests used?
- If the authors use ANOVA, are the relevant assumptions satisfied? If not, is the data transformed to satisfy them?

Domain 3: quality in presentation of results

This domain is concerned only with the quality of results shown to readers. Guidelines (see methods section of paper) advise that assessments should judge research based on results brought into the review, rather than in the original article. However, due to the nature of the topic, data synthesis is difficult and controversial. With that in mind, the best way to proceed with the currently available resources is to interpret results as they are presented in articles, but with respect to the review topic. For example, if a study is about the effect of Globalization on health outcomes, only the aspect relevant to our study (international trade Vs health outcomes) should be considered. If such a study does not present results with respect to trade Vs health despite passing inclusion criteria, then this is an issue regarding presentation of results. Primarily, this domain is aimed at detecting steering of attention to particular results and away from others. It is not always possible to detect presentation of results issues, therefore the reviewer should use her or his judgement.

- Is there any indication that any results were omitted from the study? If so, was this entirely due to space constraints or is there reason to believe that omissions were strategic?
- Are the results framed in such a way as to influence how they are perceived, their discussion or the conclusions based on them?

Domain 4: Quality from post estimation testing and analysis interpretation

This domain is concerned with testing of model robustness and other post estimation steps that are relevant to this literature. With certain methodologies such as propensity score matching, this is vital in indicating internal validity. However, there are some other issues that apply more generally. Therefore for this section, the reviewer should consider the relevant aspects, and make an informed decision on quality due to post estimation analysis and testing.

- Are all reported results considered in the inference and discussion of results?
    - Example: discussion of the confounders and exogenous effects included in the analysis.
- Is there anything else in the inference and discussion that could bias the authors' conclusion?
    - Example: citing a theoretical paper that has since been discredited to concur with their results.
    - Do you consider it a possibility that the authors have downplayed the role of a particular confounder, with statements such as "this is unlikely to affect our result". Do you agree with these statements if they are used?

Post estimation tests and further modelling

Propensity score matching

- If the matching is under 90% are various matching methodologies used to conduct sensitivity analysis?

- If the authors do conduct sensitivity analysis, does it show results to be insensitive to the matching methodology?
- Do authors use the Rosenbaum test for hidden bias? Are the results sensitive to hidden bias?

IV models with the Heckman approach

- Do the authors use the selectivity correction term? Is it significantly different from zero?

Domain 5: Other study quality issues

This domain is to capture any unique or a-typical issues in articles, and anything not covered in the other domains. If a significant amount of work is considered have a medium or low quality associated with this domain due to the same or similar issues, the tool will be revised to incorporate that issue. If no other issues are detected here, this should not influence your overall decision on the bias risk in the article. A full description of issues detected should be provided by the reviewer.

Non-trial quality assessment Fill-in sheet

Please use the following sheets to justify the decision you make for each paper considered in the quality assessment process. Please keep justifications concise, and describe only the reasoning for quality assessment.

Domain 1: Quality from the data

Domain 2: Quality associated with data approach and analysis method

Domain 3: Quality in presentation of results

Domain 4: Quality from post estimation testing and analysis interpretation

Domain 5: Other study quality issues