Appendix C: Quality assessments

1. **Cross 2009**

Non-trial quality assessment Fill-in sheet

Please use the following sheet 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.

Paper title: Does farm worker health vary between localised and globalised food supply systems?

Overall judgement: Low quality study

There were judged to be serious issues with the quality of the data used in this study, regardless of how novel, original and insightful the study was. Further, the analysis method was not applicable to this data considering the true complexity of the relationship between work and health.

Domain 1: Quality from the data

Judgement: Low Quality

Justification: The data collected was individual level HRQoL data. The sampling method was 'opportunistic' and 'non-systematic' i.e. any farm that was willing to let its workers complete the surveys used in the European sample. This was not randomised, and was highly likely to affect the validity of the results. The authors themselves acknowledge that the sample was both not representative (Top left of page 1007) and dictated by the willingness of farms to participate (further down the same paragraph). An obvious question was this: Why would a farm with lax safety policy, or worker related problems affecting the health surveys volunteer to participate in the study, regardless of whether results are anonymous or not? This raises a serious concern that any results are upwardly biased in the European sample.

Further, recruitment was not consistent in different locations. African workers were recruited on an individual level face-to-face basis by researchers from Kampala University. All of these farms were also within a specific geographical proximity of Kampala, which is not true of the recruitment strategy at other locations. It was felt that this may have affected the internal validity of the data being analysed. This was mentioned in the discussion, with the authors stating the difference in reporting across the various data collection methods. It was therefore felt that selection bias may have been a significant concern. Inclusion of data collection method, or country (as it was the same information in this case) as dummy variables to a regression would identify the significance of the data collection method to the results, and this, or any other adjustment for bias was not carried out. Therefore, there was a high risk of biased results and no attempt to address this. Overall, data should be judged to be low quality based on this issue.

Group sizes were extremely unbalanced in terms of clusters (e.g. 2 exporting farms Vs. approx. 50 non-exporting farms in Kenya). Although the two farms sampled were employing large numbers of workers (1000+), there was no indication in the paper of how many survey responses were collected from these farms. This is another quality issue due to unobserved clustering effects. If, for example 99% of data was collected at one farm and 1% the other, and the healthstate of workers was significantly different between the two, the results would lack representativeness in terms of the two locations surveyed, and this would not be reflected in the results.

Self-reported health related quality of life is known to be affected by many factors that differ from country to country. Therefore directly comparing results between countries without controlling for differences (by perhaps including country as a dummy variable to a regression analysis, or country fixed effects, or development of weights based on pilot study information) was judged to be potentially spurious. If Kenyan nationals typically assign more positive responses to the same health related issues as in other countries, then two populations with exactly the same level of health would report systematically different sets of responses to the survey. This reduces the accuracy of the results, creating another data-related limitation.

Only Kenyan data was broken down into export and non-export focused production, so the point of comparison between countries was not like-for-like. If export farms are indeed associated with better health, then comparing farms in general to them would negatively bias results, because the aggregate data would be pushed up by the export farms it contains.

Any of the above issues would individually lead to a quality assessment of medium. Therefore overall, this domain must be judged to be low quality.

Domain 2: Quality associated with data approach and analysis method

Judgement: Medium quality
Justification: The primary analysis method was non-parametric t-testing and ANOVA, and was adjusted for the distribution of the data. This did not control for confounders. Due to the complexity of HRQoL, and the unlikely situation that working conditions are the sole driver of health, it was felt that study quality in this domain was medium.
Further, a regression analysis was conducted, but this is only to seek out the role of income as a confounder. A significant relationship is identified as a result of this regression analysis. However, the primary analysis method was not adjusted to adjust for income as a confounder.
Further, it was judged to be unlikely that the only confounder to health in this situation would be income.
Overall, the analysis method was judged to be unsuitable for judgment of a complicated relationship.

Domain 3: quality in presentation of results

Judgement: High quality
Justification: The lack of a separation between export and non-export farms in the countries other than Kenya seems to be a data collection issue rather than omission from the results. The presentation of the results that they have seems to be thorough, and does not contribute to the quality as a result.

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: high quality
Justification: The conclusions are highly likely to be biased, but this is chiefly because of the data. The interpretation of results itself, in terms of the way it is conducted is thorough. Further, in the discussion, many of the weaknesses in the study are discussed, and the reader is cautioned to not read into the results too much

Domain 5: Other risks of bias

Judgement: N/A
Justification:

2. **Cross 2010**

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.

Paper title: The potential impact on farmer health of enhanced export horticultural trade between the UK and Uganda

Overall judgement: Medium quality

Similarly to the previous study by Cross et al. There were felt to be significant data related issues, but these were not judged to be serious enough to judge the whole article as low quality. However, it was felt that the results and inference in this study should be treated with caution.

Domain 1: Quality of data

Judgement: Medium quality

Justification: unlike Cross 2009, the dataset was limited to farms producing the same crops to control for differing health associations with different cultivations. Data collection for Uganda was handled by Makerere University (in Kampala), and is likely the same uganda dataset as Cross 2009. As a side note, this dataset separated exporting farms for non-exporting farms, which they did not in Cross 2009, raising a concern over the presentation of complete results.

The collection method varied between Uganda and the UK. UK farms were recruited by 'pre-existing contacts', websites and via phone directories.

Partially for confidentiality reasons (section 2.3.1), there was no indication of how many UK farms were contacted for recruitment, and thus the proportion that accepted is unknown. Eight farms were recruited into the study and workers within these had an average response rate of 56%. The subset of farms that did not agree to participate could have done so for a reason related to worker health, regardless of confidentiality being offered. For example, during a harvest period, farms may be unwilling to allow workers the time out it took to complete the data collection.

For Uganda, there was no indication that permission was sought to give questionnaires, indicating that recruitment was more on a face-to-face level. Farms had the same (or very similar) crops to the sample of UK farms, and all but one were within a 2 hour drive of Entebbe international airport. If farms close to Entebbe airport were better or worse than others in Uganda regarding worker health, the study did not represent Ugandan export worker health in general.

Differences in data collection method were still significant between the two samples, and this could have affected survey responses. However, some effort was made to remove the major issues in the earlier papers. It was felt that geographical bias could have affected the Ugandan sample, and self-selection bias could have affected the UK sample. Therefore, it was felt that we were unable to allow a score of high quality.

Overall, the quality of data was on the low-end of medium.

Domain 2: Quality of data treatment and analysis method

Judgement: Medium Quality

Justification: Since the primary analysis method was non-parametric t testing for differences in score without adjustment for confounders, it is unclear how much major confounders affected the results.

The results of both countries were compared to USA population norms. There was no way of capturing any possible effects on study results, as the authors confirm that population norms for Ugandan SF-36 do not currently exist. If the US population norms are closer to the UK than they are to the Ugandan norms, this may have made the results in one country seem more positive than they actually were.

Domain 3: Quality regarding presentation of results

Judgement: Medium quality
Justification: Table 1 shows the disaggregated (by gender) survey response averages for the UK, and their comparison to US pop norms. Table 2 shows Ugandan averages and comparison to US.

If a superscript of 'a' after the p-value was given in table 1, this meant that UK score was significantly **lower** than US population norm in table1. Superscript of 'b' meant UK score > US population norm. However, the opposite is true of table 2, where 'a' means that Ugandan score is **higher** than the US population norm. This is rather misleading and confusing. Further, although it is represented in tables 1-3, the text suggests that the results are far more conclusive than they in fact are. Many of the results are insignificant. Although the way that the results have been presented makes it difficult to read, leading to judgement of medium quality in presentation of results.

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: Medium quality
Justification: For the most part, space constraints have made the regression results difficult to read, but the inclusion of the important results in the analysis and discussion sections seems to be broadly presented. It was felt that the discussion did not raise data collection and analysis method limitations sufficiently.

The authors went on to present OLS results of health states against various explanatory variables. They show a difference in income by gender, and control for it. They also demonstrate that the income distribution is very unlikely to be normal, yet do not appear to have taken logs to reduce the effect. Further, it is demonstrated that house type, bicycle ownership and radio ownership were related to income (similarly to Cross 2009), and they are all related to health score. This would suggest that robust standard errors be used to alleviate the effect of muliticolinearity on the model. Finally, the variable 'farm' seems to describe which farm the worker's data has been collected from. However, there only appears to be one variable presented for this, meaning that the author used allocated farm id as a variable, making the coefficient meaningless (unless the data is ordered in some way). Farm id's significance suggests that farm fixed effects would be a more consistent estimator. The overall effect on the result of these issues on internal validity is difficult to gauge.

Domain 5: Other quality issues

Judgement:
Justification:

3. Kawachi 2008

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.

Paper title: Globalization and worker's health

Overall judgement: Medium quality
Overall justification: Quality issues were primarily with respect to Domain 1, but Domain 4 also raised a minor issue. Overall, the quality of the evidence was judged to be medium

Domain 1: Quality of the data

Judgement: Medium quality
Justification: For judging the situation in Korea in 2001, this would appear to be a relatively large, representative dataset. That being said, the author states that the prevalence of non-standard employment rose sharply in the late 90s. The panel data is from 2001, so as a consequence, was this a snapshot of the situation during a rapid change? If so, were the results out of date by publication? If there has been growth in non-standard employment due to more globalization, and that is associated with reporting of poor health, then was the optimal cross section in 2001?

This is a short article/letter and does not go into detail with respect to data quality. There appears to be no mention of either missing data or the representativeness of the data used. The data included 1991 male and 1378 female respondents. From the way that the logit model to calculate propensity score was described, this also provided information on age, education, household income, marital status, and occupation, type of industry and historical health status and occupational status.
As insufficient information was provided to make an informed judgement on Domain 1, it must be judged to be medium quality.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality (from the method itself. See domain 4)
Justification: The method is propensity score matching estimation using logit models to calculate the propensity scores (i.e. probability of being a precarious worker given that individual's set of descriptor variables, described above). This is a way of attempting to enable causal inference of results by removing the confounding effects of the control variables used in the matching method. If Z_i (the controls) does not contain enough information to explain the individual's propensity, then the model is subject to bias. A list of confounders included in the Z equation is given, and appears to be a reasonable selection, even including industry, job title and past health status of the individual. I would then argue that given a strong list of confounders in Z_i, the analysis method was appropriate.
The caliper width is 0.1, which seems reasonable. Individuals could be matched with others that have propensity scores up to 0.1 different from their own. Ideally this is close to 0, but that is often unfeasible considering quantity of data.

Domain 3: quality in presentation of results

Judgement: high quality
Justification: The only result presented is a graph of the odds ratios of poor self-rated health status by precarious/not precarious and male/female. The matched analysis is presented next to the unmatched, giving a fair representation of the difference made due to PSM estimation.

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: Medium quality
Justification: We were not presented with the predictive power of the propensity score model, and so we are unable to establish whether confounders were properly accounted for.

One minor point is that the authors did not conduct sensitivity analysis to test whether the matching approach selected was suitable for the sample population. Use of various matching methods is recommended, as it is difficult to establish the suitability of a matching approach.

Finally, Rosenbaum tests for hidden bias were not used. It was difficult to impossible, then, to discern internal validity issues with the model results (Again PSM was a relatively new method, so perhaps sensitivity analysis using multiple matching methodologies and Rosenbaum tests were not in common use).

The analysis interpretation includes general statements about movements towards flexible employment in neo-liberal, globalized systems (with the Caveat that EU directives have sought to reduce discrimination against part time workers since 1997). The result provided an indication that precarious work contracts increase likelihood of reporting poor health status, but the data used was from a cross section of one country.

Domain 5: Other quality issues

Judgement: N/A
Justification:

4. Bozorgmehr 2013

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.

Paper title: Trade liberalization and tuberculosis incidence: A longditudinal multi-level analysis in 22 high burden countries between 1990 and 2010

Domain 1: Quality of the data

Judgement: High quality
Justification: Additional data is available via an online appendix. There is in fact a whole section of this appendix dedicated to the presence of missing data and the approach to it. Missing data was addressed when defined as 'intermittent' (i.e. just 1 year of missing data surrounded by data) by linear interopolation. As the observation period is reasonably large (20 years), small points of interpolation to increase sample size and completeness of data are unlikely to affect the overall result. In the cases where 2 or more years were missing in a data vector, 2 years were carried forward from the last point of data, with the remaining gap being left blank. Neither of these techniques were likely to affect the result of the study.
However, the GINI coefficients were noted as having more than 80% missing data, and this was still defined as 'intermittent' in the appendix table (table 4 in the online appendix). This seemed very high considering that intermittent suggests gaps of 1 year between data points.
All countries were dropped at specific years as they all had missing data. This was considered a minor concern unless the outcome measure or explanatory variables in these years were systematically different from the years surrounding the gap.
Overall, the missing data issues are addressed very well, and the result is that the data was felt to be of a high quality

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification:
Data approach: The authors used random and fixed effects models, adjusting for time invariant and (parametrically assumed) time variant unobserved heterogeneity (separately). Further, the authors considered a wide range of confounders, assessed the quality of each one, and implemented those they assessed as most fitting. Finally, data not normally distributed was transformed accordingly. This was a reasonable data approach and worthy of high quality.
Methodology: The paper used various methods to establish a multi-level model. According to the theoretical frameworks, linear regressions were used to make links. Where non-normal distributions were detected, variables were transformed using natural logarithms. This seemed reasonable.
It would appear that the Gauss Markov assumptions were tested extensively (page 7 in paper), and adjusted for where necessary using robust standard errors, and adjustments to estimation method. As with the primary analysis method, this controlled for confounders, and included country variables. The multi-level analysis was judged as having a high quality.

The primary study method was both fixed and random effects models. The authors used a variable 'time since 1990' to account for time variant characteristics. As both random and fixed effects were used, and the appendix included various results, there was a point of comparison for readers to use in order to decide the internal validity for themselves.
Overall the approach to data and analysis method were very reasonable, and were judged to be of a high quality.

Domain 3: quality in presentation of results

Judgement: High quality

Justification: Results not presented in the paper are openly available in the online appendix. This is the most extensive presentation of results in the literature accepted into this review. There is no evidence that results were omitted.

Further, it would not appear that there is any strategic emphasis on particular results in order to influence a reader's interpretation.

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: High quality
Justification: the discussion seemed very thorough. All confounders (including the trade variables) were discussed fairly and correspondently to the results presented in tables and the appendix. The discussion of results and inference were judged as having a high quality.

Domain 5: Other risks of bias

Judgement:
Justification:

5. Gustafsson 2010

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.

Paper title: Changes in alcohol related harm in Sweden after increasing alcohol import quotas and a Danish tax decrease an interrupted time series analysis for 2000-2007

Overall judgement: High quality
Overall justification: There are some minor issues with presentation of models, and underlying problems with the data, but overall, the likelihood of the conclusion being inaccurate was judged to be low.

Domain 1: Quality of the data

Judgement: (precariously) high quality
Justification: The justification for the use of alcohol poisonings, drunk driving and violent crime assumes that proportions in comparison to alcohol consumed remain fixed as exposure increases. That is, for every unit consumed, the probability of violent crime, drunk driving or poisonings remains the same. Correlations are used to demonstrate the validity of these as means to measure alcohol abuse. It was difficult with existing data to test whether the relation between alcohol abuse and detected cases of the three indicators was linear. But, the overall effect on the conclusion of the paper was likely to be small (yet dependant on the magnitude of the change in alcohol abuse) if it is in fact not.

The hospitalisation data was judged to be internally valid and representative for Sweden, and slightly less so for crime data. This was because the authors did not give the same supporting statements to that data, and the data is not restricted to alcohol related crimes. However, a fundamental flaw was that all three indicators were vulnerable to changes in detection rates during the sample. Changes to national tax law are media issues bringing attention to issues such as alcohol abuse and violence. It would be unclear whether relevant services would react to them through striving to detect relevant crimes or negative health consequences.
Overall, it is considered that these concerns were worth raising, but unlikely to affect the overall result, unless the detection rate was affected by the policy change.

Domain 2: quality associated with data approach and analysis method

Judgement: High quality
Justification: The authors avoid biasing violent crime by removing discrete events and interpolating the preceding and following years. This was judged to be a reasonable approach to avoid skewing the distribution the crime figures. The authors also controlled for confounders using a proxy to a control group, and took logs of the harm indicators to control for non-normal distributions.

The control was data from the North of Sweden, which they argued was unlikely to be affected by the policy change. They stated just before the results section (bottom of the left column) that "It should be noted, however, that the findings for southern Sweden were little affected by the inclusion of northern Sweden as a control in the analysis and that the main results remained without this control variable (analysis not reported here)".

The authors used autoregressive integrated moving average modelling (ARIMA). Reading from Greene 2002, this is a lagged integrated model. The authors took the first difference of all the variables and included lags, then performing OLS. In order for this to be valid, data must have been stationary and should not have had any systematic seasonal differences (e.g. alcohol demand before Christmas, other festivals). To address the first issue, the authors use the Augmented Dickey-Fuller test after taking first differences and find that data are stationary. For the seasonal issue, they use the 12th monthly lag, so that each variable becomes year-on-year change.

It can be judged that the data approach and the analysis method are both robust and thus the domain scores high quality.

Domain 3: Quality of results presentation

Judgement: High quality
Justification: Results, regardless of support for the hypothesis are reported thoroughly.

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: High quality
Justification: The authors confirmed the robustness of the model in the methods section. Most of the testing required to demonstrate the internal validity of the estimation method and data was done pre-estimation and is therefore judged in domains 1 and 2.
The authors found a correlation between the tax decrease and quota removal, suggesting confounding in an analysis including both, leading them to be somewhat inconclusive in their findings. This suggests a high study quality with respect to steering of discussions and interpreting the analysis results. Overall score remains high.

Domain 5: Other risks of bias

Judgement: High quality (minor issue)
Justification: As stated by the authors, other changes that could affect detection and reporting of the three indicators would potentially bias the study, but as a judgement, the risk of this has been decided to be low.

6. Oster 2012

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.

Overall score: high quality
A large panel was used, the analysis method was reasonable and inference was also reasonable. Overall the quality of the research was high.

Domain 1: Quality with respect to data
Score: High quality
Data on incidence was predicted by UNAIDS in 2008. This covered the range 1990 to 2007. These were estimated values and HIV data is known to be difficult to accurately collect. However, the author makes this very clear to the reader, and notes that the information is based on the best available information at the time. The issue with this data is that quality related problems are unobservable to the reviewer.
Data on exports was from 3 sources. These were the World Development indicators, NBER United Nations Trade Data and Comtrade. Following methodology from a previous paper (Feenstra et al. 2004), the second two data sources are combined to create a dataset covering the period 1985-2007. This data only covers "Major exports", but the author assures the reader that this covers almost all exports in sub-saharan Africa.

The overall completeness of the data was not discussed in the paper, but was not an issue for trade and made clear through appendix table 1 for HIV. UNAIDS data existed for the whole period for all 36 countries in the panel. However, mortality based estimated data was only available for 12. As the primary analysis utilised the UNAIDS panel, this was not felt to be an issue.

The issue of trustworthiness of the data was felt to be noteworthy, but unlikely to cause serious internal validity issues in the final estimates. Section 3.1 explains the HIV data used. It is noted that until recent years, when population based testing became the favoured methodology, data on HIV was unreliable and difficult to collect. Further, UNAIDS estimates of historical prevalence of HIV are not consistent across reports. Although the author uses a single report, which provides estimates covering the period 1990 to 2007, this is likely to not be consistent with newer reports or previous. This raises questions about how accurate the UNAIDS data is, and therefore how reliable empirical data analysis based on it is.

Overall the data used was felt to be of a high quality given that the author makes it clear that the study is evidence for or against a relationship between international trade and HIV transmission, and not a demonstration of causality.

Domain 2: Quality associated with data approach and analysis method
Score: High Quality
The data did not require much manipulation in order to be used in the analysis, and the methodology used took unobserved heterogeneity across both individuals and time into account.

The control variables included were GDP per capita, country fixed effects and time fixed effects. Although this was felt to be far from exhaustive, it was not felt that the overall estimation was seriously biased by other confounders.

The estimation methodology was fixed effects. The authors used fixed effects for time as well as country, meaning that time variant characteristics were controlled for additionally to unobserved heterogeneity between individuals. The authors did not use a Hausman or alternative test, so it is not possible to know whether the assumption that the regressors were not associated with the random effect was likely to be violated, and by association whether a random effects model would be asymptotically unbiased. However, as a fixed effects model is asymptotically unbiased in the absence of endogeneity, this reflects well on the internal validity of the analysis.
Overall, the estimation methodology as well as the way that the data was treated was felt to be reasonable.

Domain 3: Quality in presentation of results
High quality
The primary results are presented in full, and (very small) graphs are provided showing the whole panel's export and HIV data for reader inspection. Overall, results were presented well and there is no reason to believe that results were strategically presented.

Domain 4: Quality in post estimation testing and analysis interpretation
Score: high quality
Two sensitivity analyses were conducted. The first was to test for level effects. The second was a set of different lagged models to test whether different lags capture more or less of the effect. This was felt to be sufficient and reasonable.
The author tests whether the process of generating incidence data undertaken was creating the relationship between exports and incidence. The author also goes onto discuss causal mechanisms, but does not claim that the evidence presented is definitive proof of causality.

Overall, it was felt that robustness testing was thorough, and discussion of the results was very good.

Domain 5: Other study quality issues

N/A

7. Alsan 2006

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.

Paper title: The effect of population health on Foreign Direct Investment Inflows to low and middle income countries

Overall score: High quality

Justification: Although the panel excluded oil dependant countries, it was judged that the data was representative with respect to the hypothesis. The methodology appeared reasonable and adjustments to the data were made to accommodate that method. Overall study quality was high.

Domain 1: Quality of the data

Judgement: High quality
Justification: the panel included 74 countries, and these were listed in the appendix. The amount of included countries appeared a representative sample of each group. However, the dataset did not cover every country. In fact, some countries were excluded based on their primary export (petroleum). It was argued that the relationship between health and FDI was biased by the oil exporters, which may be the case. For the purposes of capturing the general relationship between trade and health, it was felt that this did not significantly contribute to a risk of the study being biased. The data utilised was averaged over 10 years for each country, reducing the impact of missing data at the cost of precision. This was another minor quality issue. Overall, the quality of data was high in this study.

Domain 2: Study quality associated with data approach and analysis method

Judgement: High quality
Justification:

Data approach: unobservable heterogeneity between countries was taken into account via the methodological choice (country fixed effects). Further, time dummies were used to factor out particular events that would otherwise bias the results.

Methodological approach: Methodology was panel regression. Non-normally distributed data was logged to adjust, and the other adjustments made to the data were judged to be reasonable. Overall, the methodological selection did not reduce the quality of the study

The range of control variables included was judged to be sufficient to reduce confounding to a minimum.

Domain 3: Quality of results presentation

Judgement: High quality
Justification: Summary statistics and a correlation table were provided before the regressions to enable the reader to judge for themselves the strength of the relationships in question. The presentation of results was thorough. There was no indication that results were omitted from the study

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: High quality
Justification: The authors followed up the initial analysis by adding additional control variables to test the robustness of the results. Further, all variables were discussed in the results, and the authors stated that other variables and stratifications were used. There is no indication that the analysis interpretation reduced the study quality.

<u>Domain 5: Other risks of bias</u>

Judgement: N/A
Justification:

8. Bergh 2010

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.

Paper title: Good for living? On the relationship between globalization and life expectancy

Overall: Medium quality
Overall justification: The model is robust and the post estimation testing is also very strong. However, the omission of 24% of the panel meant that domain one was awarded a medium quality. Some testing or even a list of countries included would have addressed this issue to some extent.

Domain 1: Quality of the data

Judgement: Medium quality
Justification: The authors addressed missing data by omitting countries from the panel that had incomplete datasets. However, there was no indication of testing to see if the omission affected the results or not (e.g. testing whether averages for variables were significantly different when including the data with missing points). Without testing of this sort, and with the significant amount of countries dropped from the panel (29/121 countries dropped), the effect of this omission on estimation results was unclear. Therefore, there is no choice but to award data quality a score of medium, despite all other domains being high quality.

Domain 2: quality of data approach and analysis method

Judgement: High quality
Justification: The authors took into account both confounders and non-normally distributed data by including control variables and taking natural logarithms. There was a discussion of previous studies that established general links with growth, wealth and so on and life expectancy. They used this reasoning and previous discussion to justify the inclusion of control variables to include. The authors acknowledged 'the non-linearity between globalization and life expectancy' by taking logs of the primary input variable, the KOF (and breakdown of KOF into domains, see main text for reference and definition). Finally, dummies for both period and country were used to account for time invariant differences between countries. Overall, the data approach was strong.

The primary analysis methodology was fixed effects panel regression. Lags of the KOF and its disaggregation categories were included in the model. The authors believed that spherical errors were present and reference this approach to a monte-carlo simulation based methodological paper, citing the recommendations that those authors made. Estimations also use white-adjusted standard errors.

Finally, a sensitivity analysis was conducted by repeating the model using a random-effects model and using a different globalization index (The CGSR index). Overall, this approach is robust and caters to issues with the panel. Domain 2 was awarded high quality.

Domain 3: quality of results presentation

Judgement: High quality
Justification: The presentation of results was thorough and included full representation of the random-effects sensitivity analysis, in which various combinations of the control variables were introduced to test the reaction of the model. Very thorough and a model for other studies in this area.

Domain 4: quality of post estimation testing and analysis interpretation

Judgement: High quality
Justification: Again, the sensitivity analysis presented in table 4 demonstrated the robustness of the model. Results were similar in the baseline model with a random-effects approach. All important conclusions were mentioned in the discussion and conclusions. Results were also compared to those using a different globalization index. This was judged to be very important to conduct, as the different indices rank countries very differently (as can be seen in the paper by Zinkina, Korotayev and Andreev called 'Measuring globalization: Existing methods and their implications for teaching global studies and forecasting')

Domain 5: Other quality issues

Judgement: N/A
Justification:

9. Desbordes 2008

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.

Paper title: Public Governance, Health and Foreign Direct Investment in Sub-Saharan Africa

Overall quality: High quality
Overall justification: There are a few minor issues, but overall the quality of the study is high.

Domain 1: Quality from the data

Judgement: High quality
Justification: The authors explained that FDI was a bad proxy for multinational enterprises' actual activities immediately, but justify its use as it is the only available measure. Although results of the test are not presented to the reader, the author states that country samples of FDI data "hardly alter" when all available observations are included, even if there is missing data. It is implied that all other data except education data was complete for the sample. Education data was linearly interpolated across gaps.
The author argued that the sample was representative. They used the same reasoning as above to state that countries with missing data were unlikely to have missing data statistically significantly different from the existing data. Thus, it was an implicit argument that missing data is missing-at-random, thereby suggesting the sample is representative.
Overall, it can be argued that the quality was high. As the hypothesis regards FDI rather than Multinational Enterprise (MNE) conduct, it did not matter that FDI was a loose proxy. Finally, it was suggested that the data is complete and representative.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification: Transformations of the data were used to account for non-normal distributions. Unobserved heterogeneity was addressed through the method selection and introduction of time period dummies. Wide ranging control variables were also used to control for confounders to the issue. Random effects models were described as inferior due to fixed effects' elimination of time invariant differences between countries. However, there was no indication that a random effects model was run and compared to the fixed results to observe significant difference in coefficients (hausmann test). As consistent estimates were not tested for using this test, it became difficult to gauge the quality. However, as the primary focus was to account for between country differences, the models were shown to be robust and account for high proportions of the variation in dependant variables, the domain scored High quality.

Domain 3: quality in presentation of results

Judgement: high quality
Justification: Reporting of results is thorough and there is no evidence that any results have been omitted from the paper.

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: High quality

Justification: Presentation of a hausmann test would have been preferable, but considering the explanatory power of the models and extensive discussion of the results, this was likely a minor issue. Overall domain four scored high quality.

Domain 5: Other risks of bias

Judgement: High quality
Justification: Although it was argued that including or excluding the data with missing values did not affect regional/country calculations, the results of testing were not presented to the reader, so there was no way of knowing whether this was formally conducted or not.

10. Jorgenson 2009

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.

Paper title: foreign direct investment and the environment, the mitigating influence of institutional and civil society factors, and relationships between industrial pollution and human health: a panel study of less developed countries

Overall judgement: Medium quality
Overall justification: The problems with the data are potentially extensive, but it is very unclear to the reader. However, it is also not significantly affecting the overall conclusion. The overall study quality was medium.

Domain 1: Quality from the data

Judgement: Medium quality
Justification: The period of the panel data was 1980 to 2000. This was restricted to less developed countries. The classification of less developed was clearly presented. The issue of missing data was addressed by allowing sample sizes to vary between models. Since these (whole panel) sample sizes varied by variable significantly (49 to 519), it brought into question the representativeness and/or accuracy of the very small samples. If data was not missing at random, then important links between FDI pollution and health could have been missed, i.e. type 2 errors may have become more likely. Further, the mean number of observations per country reached as low as 2.1, and although the authors listed countries included in the sample in the appendix, they did not state the N of each country by variable or any indication of it. In isolation, the data itself could present some risk of both type 1 and 2 errors. Either a subset of the whole panel (the subset with more complete data) is over-emphasised and has a stronger connection between the variables, leading to type 1 error, or vice versa leading to type 2.
Additionally mentioned in the results section was weaknesses in the collected data. Domestic investment was for all sectors, and not just manufacturing, and there was no information on the distribution of this investment across sector. Interpretation of those results was limited to acknowledgement of their significance, which the authors comply with. Conclusions based on data where weaknesses such as these were announced were avoided throughout, and as a result this does not worsen the quality of the study.
This domain was judged as medium quality, as the extent of the problem could not be discerned, and there is no indication of how complete the panel is.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification: The dependant and independent variables were listed and discussed extensively. This included discussion and justification of manipulation. Non-normally distributed data was logged and lags were introduced to models to control for serial correlation. This was a reasonable approach.
The analysis method was split into two based on sample sizes. Small panels (defined as $n_i \leq 10$ for any i) used GLS random effects panel estimation with robust standard errors. This was justified via the value of the increased degrees of freedom in GLS estimation in comparison to other methods. As some of the countries (not specified which ones) had as little as 2 observations, the value of this was clear.
 For larger panels ($n_i > 10$), Paris-Winsten cross sectional time series was used with panel corrected standard errors. Further, in both methods, dummies for time period were used. However, dummies for country were only used in the PW models, as the other method was random effects and assumed country effects were random. To control for this, hausmann tests were used and showed that FE and RE models yielded similar results. The additional degrees of freedom from random effects was deemed superior to fixed effects estimation.
Overall, the approach to the data and the analysis method were both reasonable.

Domain 3: quality in presentation of results

Judgement: High quality
Justification: The only set of results omitted to the study was the hausmann test results, which would have required another regression table. The authors assured the reader that the difference between models was small, and this must be taken on faith. This only affects models 5 and 6, however, and is therefore unlikely to affect the overall result. Other than that, the presentation of results seems very thorough, and there is no evidence of omission based on findings.

Domain 4: Quality of post estimation testing and analysis interpretation

Judgement: High quality
Justification: As in domains 2 and 3, post estimation testing seemed thorough, although the result of hausmann tests were not presented. Quality of analysis interpretation is high as inconsistencies across models are explained well and discussion of inconveniently insignificant results was not avoided.

Domain 5: Other risks of bias

Judgement: N/A
Justification:

11. Jorgenson 2009

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.

Paper title: Political economic integration, industrial pollution and human health: A panel study of less-developed countries, 1980-2000

Domain 1: Quality of the data

Judgement: Medium quality (for the health outcome)
Justification: Extensive description of problems with data was provided. Also in the appendix was an inclusion table, showing which countries' data was included for each analysis. All countries were included for the infant mortality analysis, but for the other three hypotheses, the same countries were excluded throughout. The number of countries excluded was small. These exclusions were unlikely to affect the result of the study.

For infant mortality, the authors pointed out clearly that some of the data used was estimated based on 'census, survey or registration data'. This highlighted some vulnerability of the study to bias depending on that estimation method. Authors simply point out that 'indirect estimation' is required in non-complete vital-registration system.
The quality of the data was medium.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification: The author considered individual country effects, adjusting the methodology to account for it. Confounders were also addressed, discussed and controls introduced to all models. Models were presented without using controls and then with for the reader's comparison. The data approach was high quality.

The author used generalised least squares random effects panel models to link infant mortality to water pollution, which in turn was linked with international investment. The robustness of the model was checked through sensitivity analysis using different sets of variables. Further, the author used natural logarithms for non-normally distributed data. However, the model did not seem to take time variant characteristics into account as the other study by Jorgenson did. It was highlighted in the notes section that use of a fixed effects model had similar results (i.e. a Hausman test was conducted, suggesting the use of random effects estimation), but these were not presented. Overall, the quality of domain 2 was high.

Domain 3: quality in presentation of results

Judgement: High quality
Justification: As mentioned in domain 2, the author claimed in the notes that FE models have similar results. However, no indication of this was given. Further, note 10 stated that when including the international trade data to the infant mortality model, each investment variable became insignificant. These results were not provided in the paper however.

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: High quality
Justification: The discussion and results presented were thorough and extensive. All significant relations shown in regressions were mentioned in the text, and even the relative strength of models after including control variables was discussed. Overall, domain 4 scored very highly.

Domain 5: Other risks of bias

Judgement: N/A
Justification:

12. Levine 2006

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
High quality
The study is a cross-sectional study including spatial elements. The cross section includes a panel of 130 countries for most variables, however this is slightly less for some, falling as low as 96 for levels of Wasting. However, there is complete data from trusted sources (penn tables and World bank development indicators) for infant mortality, GDP per capita, geographical trade share and actual trade share. 4 countries were dropped from the initial 134 country sample as they all were part of the Soviet bloc and were encountering effects likely to affect all parts of the model (Bulgaria, Hungary, Poland and Romania).
The issue of measurement error is also highlighted and discussed.
Overall, the data is very old for the year the study was published (1985 for the spatial model 1990 for the main equation and 2005 for research). If the relationships between **either** spatial elements and trade or trade and child health have changed over time, this study is not estimating the relationship today.

Further, the data for the instrument and the main equation are 5 years apart. If there was a level (regardless of spatial elements) increase in international trade between 1985 and 1990 across the world, this does not affect the validity of the equation as an instrument, since the instrumental model is simply under-estimating the **coefficient** of geographical trade share, rather than the **strength of the relationship** between geography and trade share.

The quality of the data in this study is therefore judged to be high due to being mostly complete and representative. The issue of being out of date is potentially affecting how relevant the results are today, but not the validity of the actual research.

Domain 2: Quality associated with data approach and analysis method

Data treatment: The method used takes into account the difference between countries through inclusion of geographic variables to the instrumental model.

Method: The method is instrumental variables regression in the form of 2 stage least squares. The authors do not use a heckman procedure, and the instrument used is significant.
The instrumental model was found by looking up the paper originally using the instrumental equation with the same data (Frankel and Romer 1999 – Does trade cause growth?). The spatial model used gives a correlation between actual trade share and estimated trade share of 0.62, and an R^2 of 0.52 when adjusting for population and country area. As trade cannot physically affect geography, I would argue that it is exogenously generated.

Overall, I would argue that the quality of the research with respect to the methodology is high. This is due to the reasonable approach from the authors and probable exogeneity of the instrument.

Domain 3: Quality in presentation of results
High quality

A slight mistake in the text suggesting that one of the results that **was** included was not. This was the child mortality result on pages 545 and 546, for which it states the result is not provided, but it is in Table 3 (column 2). Other than that, no issues were detected with respect to presentation of results.

Domain 4: Quality from post estimation testing and analysis interpretation
High quality

There is extended discussion of instrumentation, identification of variables and using a range of different child health measures to test the sensitivity of the results. Further, there is a section discussing the causal channels between trade and child health. They acknowledge that when their models are conditioned on some variables, the effect sizes change, going on to discuss reasons for this, and the relative roles of each variable conditioned upon.
Following this, the authors conduct robustness checks. They do this by restricting the sample of countries using various criteria, finding that there is not effect on the results. Finally they condition on shocks to trade, including time variables and so on. It was felt that post estimation testing was excellent in this work.

The way that the analysis was interpreted was felt to be reasonable and fair. Their conclusions are clearly built from the results and there are no leaps in the anlalysis

Overall, domain 4 was high quality in this work

Domain 5: Other study quality issues
N/A

13. Martens 2010

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

Judgement: High quality
Justification: The primary input variable was an indexed number to indicate the level of globalization in a country. It was also an index developed by the first author of the paper. This raises some minor suspicion of a conflict of interest, however the author gives a thorough explanation of how the index is calculated, and does not overstate the ability of the index to describe the process of globalization.
There is not much discussion of missing data in this paper, despite some of the variables included reducing the amount of observations significantly. Finally, the author does not also include a more widely used globalization index such as the KOF to demonstrate external validity of the result.
However, this quality assessment is focused on the *internal* validity of studies, and aside from the issues with international representativeness of the components that make up the MGI and in some of the reduced panels (i.e. when including some of the confounding variables) no issues were detected.

Domain 2: Quality associated with data approach and analysis method

Judgement: Medium quality
Justification: there was no accounting for clustering in the study. Individual country differences were not incorporated, dummy variables for country not included and the methodology selected did not account for unobserved endogeneity. The authors extensively controlled for confounders, and presented a discussion of the confounders they did not include with justification. Further, a table of included control variables was included in the main text with official definitions, sample size and source.

It was mentioned in the methods section that the assumptions associated with both analysis methodologies were satisfied (half way down page 6, underneath table 3). This was expanded on in note 4, and the approach was reasonable.

Overall, the models they use were robust, but did not take some important factors into account. Thus, the domain could not score high quality.

Domain 3: quality in presentation of results

Judgement: medium quality **only with respect to this review**
Justification: in this review, we were interested in international trade or FDI, which were both included in the economic domain of the MGI. The model that used this domain was bivariate and therefore did not control for confounders (stated just above table 6 in the text). This raised questions about the reasons for presenting the aggregated index after controlling for confounders, and the disaggregated one before controlling.

In the context of this review, therefore, this domain must score medium quality.

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: High quality
Justification: there was extensive discussion of each final model, and some discussion of the disaggregated models using each breakdown of the MGI. Although there was some risk of affecting study results from the analysis itself, the interpretation of the results they had did not reduce quality of the study. One minor point was that different control variables were presented in the 3 tables for mortality measures. It was a little unclear whether different control variables were used in each of them.

Domain 5: Other risks of bias

Judgement: N/A
Justification:

14. Moore 2006

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.

Paper title: The health of nations in a global context: Trade, global stratification, and infant mortality rates

Overall score: High quality
Overall justification: The authors used their own classification of world system role and these were different to the literature on the topic they referred to. The quality assessment was concerned with internal validity, however. Within the study itself, quality was high.

Domain 1: Quality from the data

Judgement: High quality
Justification: The authors conducted an extensive discussion of data quality. This did not include a discussion on external validity, but focused on the study itself. The panel of countries used was large and there was also a discussion of missing or low quality data. The authors addressed this by excluding incomplete countries from the analysis and avoiding poor data sources (116 remain from the original 128 and are listed at the bottom of page 173).

Overall the quality of the study was high with respect to data.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification: The clustering of countries was on a role-level rather than a national level, and this was incorporated into the model via dummy variables. A range of control variables were also used.
The methodology was an OLS based log-log model, which accounted for non-normal distributions in the data.

Overall the methodological approach was applicable to this dataset, and the treatment of data was reasonable. The study therefore scored a high quality in this domain.

Domain 3: quality in presentation of results

Judgement: High quality
Justification: the authors were very clear when excluding variables from the analysis, and there was no evidence that any important results were omitted from the study.

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: High quality
Justification: Incremental F tests were used to indicate whether the addition of world systems role increased the R^2 of the OLS model significantly. A judgement in the study was that the insignificance of trade as a % of GDP when including world system role was an important result, but the reason for this was not thoroughly explained.

Further, as periphery 1 countries were highly dependent on international trade, trade as a % of GDP must be highly correlated with periphery 1. In order to avoid biasing the study, another model removing one or the other, using an interaction term or using an instrument to remove endogeneity would have been preferable. However, for the study they conducted, with the focus being on roles, this was unlikely to significantly bias the result. The score for quality was therefore high, since no serious limitations in relation to domain 4 were detected.

Domain 5: Other risks of bias

Judgement: high quality
Justification: The authors did not compare their own classification of world system role to other studies. Without significant knowledge of that literature, it was necessary to highlight the issue, but it was not likely to affect the results of the whole study.

15. Owen 2007

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.

Paper title: was trade good for your health?

Domain 1: Quality from the data

Judgement: High quality
Justification: The authors discussed the merits of the data they elected to utilise. However, there appeared to be a lack of discussion on data related weakness. The panel included 219 countries over the period 1960-1995 and had relatively complete data. Although data weakness was not discussed at length, the overall score for domain one was high quality.

Domain 2: Quality associated with data approach and analysis method

Judgement: High quality
Justification:

There was an emphasis on the long term relationship between trade and health because the data was at 5 year intervals.
Authors went on to describe the high variance in their key variable (openness to trade) across countries. They adjusted for this by introducing various openness measures as a sensitivity analysis.
A range of control variables were also used to account for confounding, including lagged dependant variables. Although a correlelogram was not provided, interrelationships between select variables were described descriptively. Overall the data approach was judged to be reasonable.

The statistical methodology was fixed effects estimation. The authors do not appear to have controlled for time variant characteristics in their primary analysis. However, this was following testing using lagged models to check for time variant characteristics, which revealed that their primary result was robust to them. Hausman tests were used, which indicated superiority of using fixed effects estimation also. Overall this was a model use of fixed effects estimation, with the inclusion of thorough robustness testing.

Domain 3: quality in presentation of results

Judgement: High quality
Justification: There was no evidence that any results were omitted

Domain 4: Quality from post estimation testing and analysis interpretation

Judgement: High quality
Justification: Each result in the main analysis was discussed with reference to each relevant robustness test, and each of the different trade openness measures were considered in equal weight. The inferences and conclusions weighed up all of the results without a noticeable preference or selectivity.

Considerable amounts of sensitivity analysis were included in the post estimation testing alongside testing for time-variant characteristics and reverse causality. In the one case where reverse causality was detected, this was reported and discussed in the main discussion. However, the potential endogeneity caused in the main model by the reverse-causal relationship between pharmaceutical goods, their relative price and life expectancy was not accounted for through use of an instrument or other means (e.g. attempting to omit medical imports/exports from the panel). This was judged to be of minor importance considering the extent of the robustness testing in the article.
Overall domain 4 scored high quality.

Domain 5: Other quality issues

Judgement: N/A
Justification: N/A