BIS Evaluation Summary and Peer Review

The BIS Expert Peer Review Group for Evaluation reviews all BIS impact evaluation publications, and provides an independent assessment of the methodological quality of the evaluation.

In addition to their assessment, the reviewers also provide helpful comments and suggestions for improving the clarity and reporting of the analysis. Many of the reviewers' suggestions are implemented by the authors for the final version of the publication.

The publication can be found here:
www.gov.uk/government/publications/manufacturing-advisory-service-mas-impact-analysis

Title: The Manufacturing Advisory Service - Impact Analysis
Programme evaluated: The Manufacturing Advisory Service (MAS)

Impact Evaluation Score:	**Monetisation Score:**
4 (see end of summary)	3 (see end of summary)

Time period covered by policy:	**Time period covered by evaluation:**
Jan 2012 – ongoing (Dec 2014)	Jan 2012 – Dec 2013

Contractor undertaking evaluation:	**Peer reviewers:**
BIS Analysis	Prof Maren Duvendack, London School of Economics; and Dr Henry Overman, University of East Anglia

Type of evaluation:
Quasi-experimental impact analysis

Description of policy/programme and rationale for intervention:

SMEs can be poorly equipped to grow due to lack of expertise in the latest manufacturing techniques and processes. They often find it difficult to access expert manufacturing advice and it is this that stops them from realising the productivity benefits of improved manufacturing techniques. MAS is the only support scheme in the country that provides expert manufacturing help. Manufacturers can take advantage of a free manufacturing review by a MAS advisor to identify key priority areas or access funding for improvement projects to increase efficiency, develop new products and boost sales. MAS acts in support of Industrial Strategy by:

- Positioning suppliers to take advantage of inward and new infrastructure investment as well as building future capacity.
- Strengthening underperforming supply chains by helping SMEs meet prime requirements on quality, cost and delivery times.
- Helping suppliers respond to external shocks such as new regulations or site closures.
- Protecting jobs, creating new jobs and enabling growth in gross-value added (GVA).

Summary of key evaluation findings:

This report is an analytical paper, rather than a full evaluation. We present a methodological framework for evaluating MAS that is an improvement on that used in previous evaluations, as we have been able to identify a counterfactual, estimate the scale of selection biases and avoid using self-forecast or self-reported growth. We are able to estimate the average GVA benefit per business over the treatment period - £15,000 - £30,000 – but avoid explicitly calculating an overall benefit-cost-ratio for the scheme due to a lack of evidence on additionality. Despite the improvements associated with this methodology, we describe some significant uncertainties and methodological issues that should be addressed in future.

Summary of cost-benefit/cost-effectiveness analysis (if applicable):

N/A

Policy response to the evaluation:

The suggested changes to the matching methodology will be investigated as part of future evaluation studies.

Evaluation methodology

Description of methodology:

The methodology uses the following techniques:

- ***Data linking*** – We have linked the database of MAS clients to a comprehensive ONS database, covering variables such as turnover, employees and sector over the period 2010 to 2013. This has allowed us to find a comparator group of businesses that have not received MAS support and estimate a measure of each business's Gross Value Added (GVA) both *before* and *after* they have received MAS support.
- ***Matching Methods*** – We have used matching methods to identify a group of businesses not in receipt of support (the "control group"), matched to a group of MAS clients (the "treatment group") on key parameters, such as sector, initial turnover and business birthdate. By assuming these parameters are important in determining economic performance and likelihood of receiving support, we can use this matched group as a counterfactual to assess the economic impact.
- ***Difference-in-difference (DiD)*** – The impact of MAS support can then be estimated by calculating the difference in GVA growth over the period of interest between the treatment group and the control group.

We then use a series of different quasi-experiments to estimate the additional GVA growth associated with self-selection bias. Finally we estimate the scale of additional GVA growth associated with the selection bias introduced by MAS advisors when choosing businesses for grant funding. As there is no robust evidence available to assess advisor-selection bias we use a wide range of values. We avoid extrapolating these results up to estimate the aggregate economic benefit of MAS as we lack the data to estimate the different components of additionality.

Does the evaluation review the published policy objectives?

The analysis does review the policy objectives but many of the intended impacts, such as business strategy, innovation and efficiency are not easily measureable. We focus on GVA growth as an indirect, but measureable impact of the policy. We also briefly cover employment growth.

At what level are the main intended outputs and/or outcomes expected to occur? (What is the unit of analysis? For example: universities, businesses, individuals or nationally)

At a business level.

Has sufficient time lapsed for the initial/full benefits to be estimated?

The analysis suggests that sufficient time has lapsed for initial benefits to be observable but there may be additional benefits over subsequent years.

Peer review

Comments on the appropriateness of data and outcomes:

Maren Duvendack:

I should note that I have been in touch with the author during the course of this evaluation to comment on the interim report. We also had phone conversations discussing the approach to matching.

When I commented on the interim report I raised the following key issues:

- *Timing of the evaluation – is it realistic to expect impacts to emerge after 1 year only?*
- *Contamination of control groups – how valid are they?*
- *Concerns about the success of the data linking exercise*
- *Serious concerns about selection bias issues, selection either by self or by the MAS advisor*
- *Other interventions taking place at the same time of MAS which might be complementary or in fact drive the impact that is observed*
- *Comments on the analytical approach and some of the results presented, as well as issues related to the outcome variable that has been used and the different experiments that have been conducted*

The author has addressed most of these comments in the final report but some of my concerns regarding timing of the evaluation and contamination of control groups still hold. These two issues are linked to limitations of the data sources that are available and the author has been very critical and honest about these limitations. Thus, not much that can be done about this at this stage but this leads me to raise the issue of aiming to improve data sources and making them more relevant and appropriate for such an evaluation exercise.

Linked to this is the problem of having only one decent outcome variable – GVA – to assess the economic impact of MAS. MAS also has employment related objectives which are not captured by this evaluation, again, due to data limitations as far as I understand this. I will return to the issue of data limitations below when I comment on external validity.

Henry Overman:

The report uses admin data (MAS monitoring) matched to a secondary data source (the Inter-Departmental Business Register). The IDBR provides information on employment and turnover as well as providing detail on the sector of the firm. The report translates turnover in to gross value added based on the four-digit sector average ratio and then conducts the analysis mostly for turnover. As per my comments on the interim report, I would have preferred for the results to be reported in terms of the observed variables (turnover). If cost-effectiveness discussion required a conversion to output, I would rather this have been done after the analysis was conducted. I do not think this would make much difference to the results – but it is a cleaner way to proceed. Aside from this issue, I do not have any other major concerns. Data linking rates (of IDBR to MAS monitoring data) appear broadly satisfactory. There is some bias towards larger firms in the matched data – as reported in tables 12 and 13 on p.24. This is to be expected and unlikely to present a major problem for the subsequent analysis.

Comments on internal validity:

Maren Duvendack:

When reviewing the interim report it became apparent that it is highly likely that a number of biases will drive impact estimates. It became also clear that the methodological approach proposed here would not be able to fully address these biases. Thus, doubts remain about how well internal validity has been addressed. Section 1.7 is useful in this regard as it describes the biases that are likely to threaten internal validity.

Section 2 which sets out the methodological approach is a considerable improvement from the methodology section presented in the interim report. Section 2 in the final report provides a better explanation of the data linking exercise as well as the matching approach.

Having said that, coming from a methodological background with a key interest in matching methods I would have liked to see a more in depth discussion of the various matching approaches that have been tested: PSM vs CEM vs NNM. From an academic perspective I think it would be quite interesting to explore this further but I understand that this is not the focus of the report.

When employing PSM one can check for quality of the matches as well as conduct sensitivity analysis to deal with selection due to unobservables. I understand that this is not possible to the same extent with CEM and NNM which is a bit of a drawback. But I know that CEM provides a matching summary at the end breaking down the sample into matched and unmatched observations, this gives some sort of indication of the matching quality. I am not sure whether NNM reports something similar as I haven't used NNM very much but if so then this could or should be reported in the appendix. I think it is important to engage with the quality of the matches as this adds to the credibility of the approach that has been taken. Regarding sensitivity analysis (as set out by Rosenbaum, 2002), I don't think that has been operationalised in the context of either CEM or NNM yet, so the role of the unobservables cannot be quantified as a result which makes triangulation with the qualitative evidence even more important.

It is excellent that the STATA code has been published in the appendix. Though an estimation dataset as well as the code should be made available for download. There is the UK data archive for example but there might be other archives too. Given the renewed interest in replication and data sharing this is an important topic. E.g. many journals and funding bodies (e.g. ESRC) now have data sharing policies in place where it is mandatory to provide data as well as code. See for example the American Economic Review (AER), Econometrica, the Journal of Applied Econometrics (JAE) and others. The AER argues that

> *"For econometric and simulation papers, the minimum requirement should include the data set(s) and programs used to run the final models, plus a description of how previous intermediate data sets and programs were employed to create the final data set(s)".*[1]

[1] http://www.aeaweb.org/aer/data.php, accessed 5 August 2014.

This requirement is "an important step towards a more transparent and credible applied economic research" (Palmer-Jones and Camfield, 2013:1610). In fact, we should go further than this and authors should not just provide datasets and code that run the final results published in a report/paper but also provide a description of how the raw data was managed and compiled into a final estimation dataset as a lot of data manipulation can take place between original (raw) and final data sets that is not carefully documented (Palmer-Jones and Camfield, 2013). This is an important point especially in this particular case as the data linking exercise was complex and rather challenging.

Henry Overman:

I have a number of concerns with regard to the internal validity of the estimates. The first concerns the fact that the matching procedure does not appear to be producing a very good match. For example, in figure 10, simply extrapolating the pre-treatment trends for control and treated observed in 2010 and 2011 appears to give a GVA gap in 2013 which is similar in magnitude to that actually observed post-treatment. Figure 15 provides an even more striking example for experiment B (the matching of L2s to No MAS) – here the 2011 (i.e. pre-treatment) GVA difference is already statistically significant. Figures 16 and 17 demonstrate the same problem for employment (again for experiment B). Second, and related, it is striking that the main impact of treatment appears to arise because growth rate suddenly drops in the non-treated group (rather than the treated group showing an increase relative to trend). See, for example, figure 10. Some of this could be due to the striking differences in death rates – however here it's impossible to assess the differences in pre-trends because the results are only reported based on 2011 matching (so firms in both treatment and control need to exist at that point). It would be reassuring if matching on the same variables on the basis of 2010 data did not produce significant differences in pre-treatment trends. The report does not consider this possibility. Similar concerns apply to experiment D (i.e. there appears to be a positive difference in pre-trends). For experiment C, the pre-trend comparison appears to work in the opposite direction and is likely to biased downwards the findings (although treatment has no significant effect in these estimates). To be convincing, the report really needs more discussion of these issues.

Comments on external validity:

Maren Duvendack:

External validity is addressed by the numerous experiments that are conducted. I think this is crucial as MAS support comes in many different forms. The different levels of MAS support can be disentangled, at least to a certain degree, by running different sub-group comparisons (called experiments here). A number of experiments have been presented in this final report.

The following is something I draw attention to in most of my BIS evaluation reviews: The importance of presenting a strong theory of change. Theory-based approaches to evaluation experience increased attention; many claim that a theory-based approach can strengthen claims to external validity. The logic model set out in section 1.4 is a bit thin in my view and does not fully capture how MAS actually works. MAS is a complex intervention with different levels of support and interconnected objectives and targets which should be better captured in the logic model. This evaluation could have benefitted from a more elaborate theory of change.

A theory of change here could have identified the links between the different levels of MAS support and how those lead to the desired overall outcomes of MAS. The following references provide a good introduction to exploring theory-based approaches to evaluation: Funnell and Rogers, 2011; Rogers, 2008.

A stronger Theory of Change could have also helped to derive better and more relevant indicators for the MAS monitoring database. When reviewing the MAS interim report I was surprised that with all these data sources available, only the GVA could be generated as a relevant outcome variable. MAS' objectives go beyond GVA and in fact this evaluation is concerned with examining the economic impacts of MAS; the GVA can hardly represent the whole spectrum of economic impacts. My understanding is that data limitations and challenges of successfully linking data led to this problem. Either way, the point here is that with a better theory of change indicators more closely aligned with the MAS objectives could have been derived which in turn could have been captured by the MAS monitoring database. A richer monitoring database could be beneficial for subsequent MAS evaluations.

This report is critical of the methodological approach adopted, quite rightly so, but it is not the analytical approach taken here that is at fault, the data sources used are the problem. The famous computing principle of garbage in, garbage springs to mind here. Just because you can employ sophisticated econometric techniques does not mean it makes sense to do so when the underlying data sources are not appropriate. So I think it is important to review existing data sources and think about what sort of data should be collected in the future to make evaluations of programmes like MAS easier.

References:

Funnell, S., & Rogers, P. 2011. Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. San Francisco: Wiley & Sons.

Rogers, P.J. 2008. "Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions." Evaluation, 14(1), pp.29-48.

Henry Overman:

Assuming that the purpose of the evaluation is simply to test the effectiveness of MAS provision for manufacturing firms, then no there are no obvious question of external validity. Because there is selection in to treatment, the major concern relates to internal validity as discussed above.

Comments on the quality of inferences and establishing causation:

Maren Duvendack:

The conclusion and recommendation section on page 9 as well as section 4.6 quite nicely summarise the concerns with regard to the methodological approach taken. In principle matching approaches and other econometric techniques aim to establish causation beyond reasonable doubt but given the complexities encountered in this particular case we cannot fully attribute the impacts we observe to the MAS programme.

The author has done the best he could given data limitations and challenges related to data linking, he has frankly set out the issues related to selection bias, and so on, and made some good suggestions for improving the methodological approach. I agree with his assessment that the analysis should not be used for making specific policy recommendations as it still has many uncertainties.

Henry Overman:

My prior was that establishing causality would be most challenging for experiments A and B – which needed to construct non-MAS control groups. Because there is selection in to treatment (firms have to first approach the MAS service before being considered for treatment) firms that receive MAS treatment (of some kind) might be expected to be different from firms that do not. Experiment D suffers from the same problem, but only if the timing of selection in to treatment is driven by firm specific factors. In some senses this is a weaker assumption – although in practice the pre-treatment problem identified above applies in all three situations. Experiment C suffers from a different selection problem – that advisors may choose L4 and L2 and likely do so on the basis of factors that are unobservable from the IDBR or MAS data. Again, the discussion of pre-trends suggests that this is an issue in practice. As a result, it is hard to assess the extent to which the report has been able to establish the causal impact of the programme.

There is very little discussion of the treatment of standard errors and the report should be clearer on this issue.

Any other comments:

Maren Duvendack:

Just a couple of comments related to the presentational nature of this report:

- *A table with abbreviations might be useful as abbreviations that are not explained in the text crop up occasionally.*
- *The executive summary reports the impact estimates of £90,000 and £15,000-30,000 – it is not clear whether these are statistically significant, one has to read the full report to work this out. So I would make this clear in the summary already.*
- *Annex 3 could do with a bit more annotation, at least headers to the graphs and a brief description would be helpful.*

I thought that the qualitative section (section 1.6) was useful to provide more contextual background and allow triangulation. However, in section 4 where results are discussed the author then does not refer back to the qualitative evidence and triangulate it with the econometric results which I think is a bit of a shame. In the initial part of the report I got the impression that the qualitative evidence was specifically collected for triangulation purposes, thus in the interpretation section I would have expected more of a discussion of this.

A final note on the recommendation of using a randomised control trial, it is important to note that RCTs usually capture short term impacts unless a follow-up is planned after a few years. RCTs are also expensive and time consuming requiring massive buy-in from the programme implementer, thus I would carefully think about the usefulness of an RCT in this context before embarking on one. Page 60 raises some concerns in this regard.

Henry Overman:

Overall I think the report demonstrates the potential of matched data to help further our understanding of the impact of this kind of treatment. The problems of selection in to treatment (or type of treatment) would also apply to more traditional approaches based on be-spoke survey data. The high matching rate suggests that secondary data provides a viable alternative to be-spoke data – providing large sample sizes (although clearly involving a trade-off in terms of outcome coverage – see my comments above on turnover). An additional advantage of secondary data – as demonstrated here – is the ability to identify differences in pre-trends and thus assess the extent to which this is driving any identified treatment effects. The methods applied are certainly an improvement on self-reported additionality and are an important step towards developing a robust methodology for evaluating impact of MAS.

The report does less well in terms of its objective to identify the impact of MAS and the effectiveness of different types of support. To some degree, this is a problem of implementation. The style of write-up and remaining concerns over internal validity should have been addressed. The department should reflect on why this did not happen (capacity, time available, degree of support for the analysis?). The second problem is more fundamental and reflects the fact that this is a well-established scheme where firms select in to treatment. If the department feels that the scheme must offer treatment to all firms that approach the service, then establishing the overall effect of the programme will be challenging.

Improving on the analysis developed on this report might provide more confidence on the overall effect of the programme, but concerns over selection on unobservables will remain. There is more scope for addressing questions concerning the cost effectiveness of different types of firm support – although properly assessing these would require some randomisation in to treatment type. Again, improving the analysis developed in this report would provide more confidence on the effect of specific treatment, but concerns over selection in to treatment type on the basis of unobservables would remain. Some of these issues are covered in the final section of the report and I would urge the Department to carefully consider them.

Cost-effectiveness and cost-benefit summary

Justification for monetisation score:

To date, no assessment has been made of additionality effects for MAS and so no BCR has been estimated for the program. A survey of grant recipients has now been established for MAS clients, which will help to estimate additionality effects. This uses an approach which is consistent with the approach for Growth Accelerator (GA) – MAS and GA have now been combined as a new Business Growth Service.

Sensitivity analysis/key assumptions:

Direct costs to Exchequer of programme:

£m	Total	Year 0	Year 1	Year 2
Total				

Economic costs and benefits of programme:

Price base year		Present value base year		Discount rate	

	Costs (£m)			Benefits (£m)				
	Transition (constant price)	Average annual	Total (PV)	Transition (constant price)	Average annual	Total (PV)	NPV (£m)	Net BCR[2]
Low								
Best estimate								
High								

[2] PV of net benefits / PV of net costs

Description and size of key monetised costs:
Other key non-monetised costs:
Description and size of key monetised benefits:
Other key non-monetised benefits:
Robustness of monetised costs and benefits:
The costs included in the report have been provided by Grant Thornton who administer the MAS program and cover administrative and grant expenditure from administrative sources. The estimated average GVA benefit per business over the treatment period of between **£15 - £30,000** is provisional. As noted in the report, there are timing issues with updating of IDBR data and an evaluation of economic benefits is normally undertaken over a longer time-scale following an intervention, typically over 2-3 years for a scheme like MAS.

Peer Review

Evaluation peer review comments on comprehensiveness, clarity, robustness and best practice of cost benefit/cost effectiveness analysis:

Henry Overman:

Given my concerns above (including in the way that turnover is translated in to output) I would not view the £15,000-£30,000 GVA figures provided on p. 61 as a causal estimate of programme impact.

I would also like to see some discussion of costs to the firm of programme participation (time costs, etc).

The report does not undertake any further cost benefit analysis so I cannot comment further (p. 62)

Note on Impact Evaluation and Monetisation Scores

Impact Evaluation Score

The higher the score the more capable the evaluations are to demonstrate that the *outcome observed is due to or caused by the intervention*. Impact scale follows new guidance on 'Quality on Impact Evaluation' which has been approved by the Cross Government Evaluation Group and will be published alongside the Magenta Book.

- Score 5: Random allocation of treatment and control group or matched treatment and control group. Actual before and after data in both groups.
- Score 4: Treatment and comparison group. Actual before and after data in both groups.
- Score 3: Predicted versus actual (modelled), predicted based on actual baseline data.
- Score 2: Actual before and after
- Score 1: No baseline data

Monetisation Score

The higher the score the more information the evaluation contains in terms of *analysing the cost* of the intervention and the additional *benefits* to the economy.

- Score 5: Input, output, outcome data additional Benefit Cost Ratio (BCR), NPV set aside some other not monetised impact measures, fuller cost benefit analysis or cost effectiveness analysis that compares the costs of alternative ways of producing the same or similar outputs
- Score 4: Input, output, outcome data, calculation of additional Benefit Cost Ratio, Net Present Value
- Score 3: Input, output, outcome data calculation of Gross BCR not additional or not clear if additional
- Score 2: Gross BCR not available, as either input or output data are not available
- Score 1: No monetisation at all