



REMAS

Analysis of initial sample data for the United Kingdom

Roger Salmons
Environment Group
Policy Studies Institute

October 2004

Introduction

This report presents the results of a detailed statistical analysis of the relationship between EMS classification, operator performance and environmental outcomes, for a sample of fifty-seven sites, operating across ten industrial sectors in the United Kingdom.

The specific objectives of the analysis are:

- to assess whether the implementation of an accredited environmental management system (EMS) leads to improved operator performance (as measured by a site's EMA scores);
- to assess whether higher levels of operator performance lead to better environmental outcomes (as measured by a site's EP scores); and if so, to identify which dimensions of operator performance have the greatest impact.

The report is in four sections. The first section provides a short overview of the main findings of the analysis and the recommended actions. In the second section, the underlying framework that has been adopted for the analysis is outlined, the sample data reviewed, and the results summarised. A number of potential issues that might have affected the analysis are discussed in section three. The final section comprises a number of appendices, which provide a detailed description of the analysis, together with a brief introduction to the topic of regression analysis and hypothesis testing.



1. Overview

Key findings

- The sample provides strong evidence to support the hypothesis that the adoption of an accredited environmental management system leads to an overall improvement in operator performance (as measured by a site's average EMA score); with EMAS having a greater beneficial impact than ISO 14001.
- The same is true for seven of the nine individual dimensions of operator performance. However, the sample provides no evidence that the adoption of ISO 14001 leads to an improvement in a site's *commitment to training and awareness* (EMA13), or that the adoption of EMAS leads to an incremental improvement over ISO 14001 for *operational and risk management* (EMA12).
- Allowing for differences between industrial sectors, the sample provides no evidence to support the hypothesis that higher levels of operator performance lead to better environmental outcomes in terms of *compliance* (EP1) and *conduct* (EP2). However, this may be due to the methodology used to construct the scores for these measures.
- In terms of process efficiency (EP3-EP5) and releases (EP6-EP8), there is some support for the hypothesis that higher levels of operator performance lead to better environmental outcomes – at least for some sectors. However, the evidence is not very strong.
- For the individual dimensions of operator performance, the sample provides no evidence to support the hypothesis that either *training and awareness* (EMA13), or *documentation control* (EMA17) have a positive impact on any of the environmental outcomes.
- The high level of correlation between the other seven dimensions of operator performance, makes it difficult to draw definitive conclusions about their relative impacts. However, the sample provides some tentative evidence to suggest that the impacts differ between the different types of environmental outcome.
- For process efficiency (EP3 and EP5), *environmental policy* (EMA11) and *compliance and conformance control* (EMA14) appear to have the greatest impact.
- For releases (EP6 and EP7), *performance monitoring* (EMA15), *environmental reporting* (EMA19) and *compliance and conformance control* (EMA14) appear to have the greatest impact
- However, these tentative conclusions should be treated with a considerable degree of caution, as the results on which they are based may be very sensitive to small changes in the sample composition.



Recommendations

- Increase the sample size for each EP measure to a minimum of around 100 sites – either by increasing the total sample size, or by increasing the sample penetration for each measure, or both.
- Increase the variability of the EMA scores for each dimension of operator performance – by increasing the proportions of sample sites with EMAS, and with no EMS.
- Review the methodology that is used to calculate the *compliance* measure (EP1), to ensure that it properly reflects the “concept” of regulatory compliance
- Review the definition of the indicator for unresolved complaints that is used in the calculation of the *conduct* measure (EP2), and (possibly) the methodology that is used for combining the two component indicators
- Review the correlation between the indicator scores for each process efficiency measure (EP3 – EP5) and each releases measure (EP6 – EP8), to check the validity of the treatment of “missing” indicator values.
- Review the cut-off values that are used to determine the scores for the indicators used in the calculation of each process efficiency measure (EP3 – EP5) and each releases measure (EP6 – EP8), to check that the implied “equivalence contours” between indicators are correct.



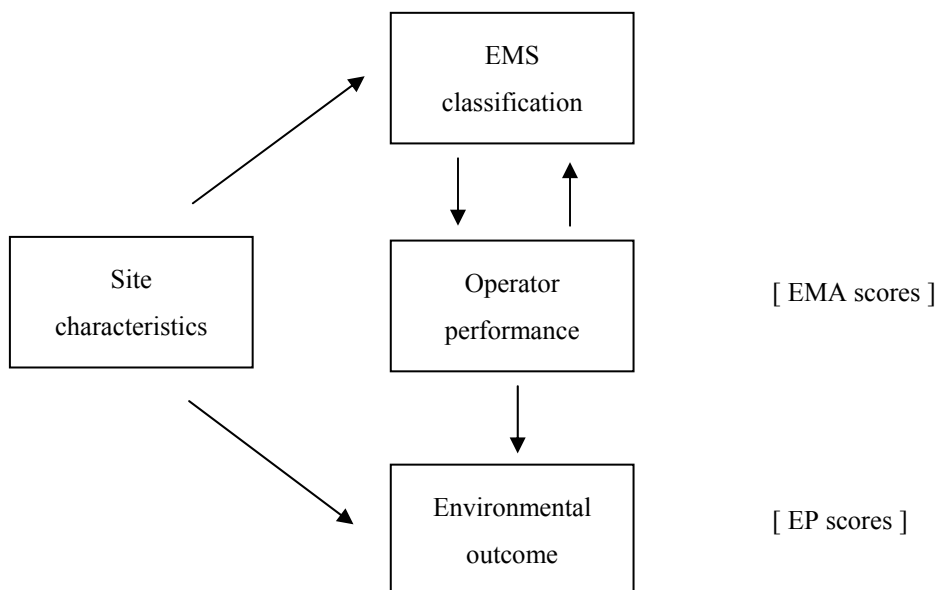
2. Summary of analysis

2.1 Framework for the analysis

Figure 2.1 provides a schematic representation of the model that underpins the analysis. The model is based on the following three assumptions (or hypotheses):

- that the adoption of a formal (accredited) environmental management system (EMS) leads to higher levels of operator performance – as measured by a site’s EMA scores;
- that higher levels of operator performance lead to better environmental outcomes – as measured by a site’s EP scores;
- that site characteristics (such as size and industrial sector) may influence its EMS classification, and / or the environmental outcomes for a given level of operator performance.

Figure 2.1: Model

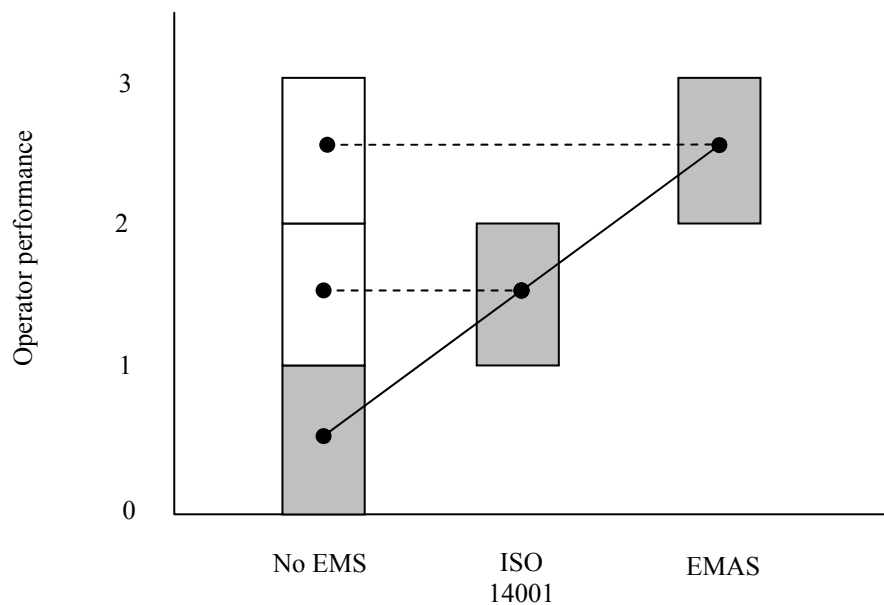


The model also allows for the possibility that sites with higher levels of operator performance in the absence of a formal EMS may be more likely to seek accreditation. This correlation may reflect underlying management attitudes towards the importance of environmental management, or lower costs of meeting the necessary level of performance for the accredited schemes. The importance of taking this “endogeneity” into account is illustrated in Figure 2.2, which shows the ranges of operator performance scores for the three EMS classifications for a hypothetical example.



For each EMS class, the grey bar gives the observed range of performance scores for each EMS class; while the two white bars for the “No EMS” class represent the ranges of scores that would have been observed for sites with ISO 14001 and EMAS if they had not implemented these systems. Thus, the sites that have adopted EMAS would have formed the top third of the distribution if all sites operated without a formal EMS; while those that have adopted ISO 14001 would have formed the middle third. Based on the observed scores it might appear that the adoption of ISO 14001 improves the average score from 0.5 to 1.5, and that upgrading to EMAS further improves the average score to 2.5. However, it is clear that such a conclusion would be incorrect. In reality, the average scores of the sites that have adopted these systems are exactly the same as they would have been if they had not.

Figure 2.2: EMS classification and operator performance



Reflecting the structure of the model, the analysis of the sample data is broken down into two separate stages. In the first stage, the relationship between EMS classification and operator performance is analysed, and the “causal” impacts of ISO 14001 and EMAS on the different dimensions of operator performance are estimated. In the second stage, the impact of operator performance on the different dimensions of environmental outcome is assessed. However, before considering the results of these analyses, it is instructive to review the sample data.

2.2 Review of the sample data

a) Sample structure



The sample comprises fifty-seven sites operating in the United Kingdom. Table 2.1 gives a breakdown of the sample by industrial sector and EMS classification.¹ As can be seen, there are considerable differences between the classification profiles of the sectors. In particular, the cement, combustion and organic chemicals sectors are overweight in sites with EMAS, ISO 14001 and No EMS respectively.

Table 2.1: Breakdown of sample by industrial sector / EMS classification

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	8	3	2	13
Combustion	0	8	1	9
Inorganic chemicals	0	4	2	6
Organic chemicals	1	5	6	12
Paper & pulp	0	3	2	5
Other	0	8	4	12
Total	9	31	17	57

Table 2.2 gives a breakdown of the sample by plant size (as measured by number of plant employees) and company size (as measured by number of plants). In terms of these two dimensions, the sample is more balanced. However, the sample is slightly overweight in small sites (1-100 employees) operated by small companies (1-10 sites), and in large sites (> 100 employees) operated by large companies (> 10 sites); reflecting a weak correlation between the two characteristics.

Table 2.2: Breakdown of sample by plant size / company size

Plant size (number of employees)	Company size (number of plants)				Total
	1 -3	4 - 10	11 - 20	20 +	
1 - 10	1	0	0	0	1
11 - 50	9	3	2	1	15
51 - 100	1	4	2	0	7
101 - 250	4	1	8	0	13
251 - 500	9	0	3	2	14
500 +	0	2	2	3	7
Total	24	10	17	6	57

b) Operator performance

¹ The “Other” sector comprises sites from the Aluminium, Food, Refining, Glass and Other Minerals sectors.



Operator performance is measured on an index scale – ranging from a minimum value of one, to a maximum value of two.² For each site, a score has been provided for overall performance (EMA), and for performance along each of nine individual dimensions of environmental management (EMA11 – EMA19). These are listed in Table 2.3.

The scores are based on the site’s responses to fifty-six multiple-choice questions, which have been subjected to a “quality control” checks by the Environment Agency. For each question, the site is scored on a five-point scale (0 – 4), with a higher score corresponding to better performance. The score for each dimension is then calculated as the simple average of the scores for the questions relating to that dimension, while the overall score is calculated as the simple average of the scores for all fifty-six questions. The raw scores are then scaled so that a score of 0 is set equal to 1, and a score of 4 is set equal to 2. The number of questions relating to each dimension is given in Table 2.3, and a detailed breakdown of the allocation of questions to the dimensions is provided in Appendix A.³

Table 2.3: Dimensions of operator performance

Measure	Description	No. of questions
EMA11	Environmental policy	4
EMA12	Operational and risk management	13
EMA13	Commitment to training and awareness	4
EMA14	Compliance and conformance control	8
EMA15	Performance monitoring	13
EMA16	Open communication culture	7
EMA17	Documentation control	2
EMA18	Management review	5
EMA19	Reporting environmental performance	6
EMA	Overall	56

Table 2.4 provides some basic summary statistics for the different dimensions of operator performance. For all nine dimensions (plus the overall average), scores have been provided for all fifty-seven of the sample sites. The mean score differs between dimensions, ranging from 1.52 for *open communication culture* (EMA16) to 1.74 for *environmental policy* (EMA11). The variability of the scores also differs between dimensions; with the standard deviation of the *management review* scores (EMA18) and the *environmental reporting* scores (EMA19) being almost twice as great as that of the *performance monitoring* scores (EMA15).

² The minimum value is set equal to one in order to ensure that the log of the index value is always defined. The choice of the minimum and maximum values for the scale is completely arbitrary. It does not affect the conclusions of the analysis, although it does affect the values of the constant term and the coefficients of the explanatory variables in the regression equations.

³ It should be noted that seven of the questions apply to more than one dimension. Consequently, the overall score is only approximately equal to the weighted average of the scores for the individual dimensions – where the weights are based on the respective number of questions.



Table 2.4: Summary statistics for operator performance measures

Measure	No. of sites	Mean	Minimum	Maximum	Standard deviation	Spread
EMA	57	1.66	1.34	1.90	0.15	0.56
EMA11	57	1.74	1.35	2.00	0.16	0.65
EMA12	57	1.67	1.13	1.97	0.16	0.84
EMA13	57	1.63	1.13	1.94	0.22	0.81
EMA14	57	1.68	1.08	2.00	0.21	0.92
EMA15	57	1.66	1.30	1.94	0.15	0.64
EMA16	57	1.52	1.00	1.93	0.24	0.93
EMA17	57	1.65	1.25	2.00	0.19	0.75
EMA18	57	1.64	1.00	2.00	0.25	1.00
EMA19	57	1.57	1.00	2.00	0.25	1.00

The pair-wise correlation coefficients shown in Table 2.5 provide an indication of the degree of correlation between the scores along the respective dimensions. A positive value for the coefficient indicates that the scores are directly related; a negative value, that they are inversely related (i.e. a high score for one measure is associated with a low score for the other). The magnitude of the coefficient provides an indication of the strength of the association. The closer the magnitude is to one, the stronger the correlation; the closer to zero, the weaker the correlation. An arbitrary cut-off value of 0.6 has been used to identify the pairs of measures with relatively high correlation, which are highlighted in grey.

Table 2.5: Correlation matrix for EMA measures

	EMA11	EMA12	EMA13	EMA14	EMA15	EMA16	EMA17	EMA18	EMA19
EMA11	1.00								
EMA12	0.60	1.00							
EMA13	0.44	0.29	1.00						
EMA14	0.78	0.70	0.37	1.00					
EMA15	0.59	0.54	0.38	0.63	1.00				
EMA16	0.58	0.53	0.30	0.65	0.66	1.00			
EMA17	0.44	0.47	0.16	0.52	0.32	0.45	1.00		
EMA18	0.82	0.60	0.41	0.84	0.54	0.67	0.49	1.00	
EMA19	0.60	0.53	0.37	0.61	0.63	0.72	0.24	0.64	1.00



It is clear that – with the exception of *commitment to training and awareness* (EMA13) and *documentation control* (EMA17) – there is a relatively strong positive association between the different dimensions of operator performance. The magnitudes of the correlation coefficients between the other seven dimensions range from 0.53 to 0.84, with the majority exceeding 0.6. This, of course, is not very surprising. If a site performs well along one dimension, it might be expected to perform well along the others. Indeed, the relatively low correlation coefficients for EMA13 and EMA17 are more interesting. However, as will be discussed later, the strong correlation between the measures makes it difficult to isolate their individual impacts on the environmental outcomes.

c) Environmental outcomes

The environmental outcomes are also measured on an index scale – with a minimum value of one, and a maximum value of two. Eight different outcomes have been assessed, although for most sites, scores are only available for a subset of the total. The outcomes are listed in Table 2.6: two relate to regulation (EP1 – EP2); three relate to process efficiency (EP3 – EP5); and three relate to releases (EP6 – EP8).

Table 2.6: Environmental outcome variables

Measure	Description	No. of indicators
EP1	Compliance	5
EP2	Conduct	2
EP3	Process efficiency – raw materials	1
EP4	Process efficiency - water	1 - 2
EP5	Process efficiency - energy	1 - 3
EP6	Releases - air	1 - 3
EP7	Releases - water	1 - 3
EP8	Releases - land	1 - 3

With the exception of the process efficiency measure for raw materials (EP3), all of the measures are calculated as the simple average of the index scores for a number of sector-specific indicators; which in turn are derived from actual values provided by the sites. The number of indicators included in the calculation for each outcome measure is shown in Table 2.6. For two of the process efficiency measures (EP4 – EP5), and all of the releases measures (EP6 – EP8), sites could choose from a list which indicators they provided information for, up to a stated maximum number.⁴ Consequently, for these measures, the number of indicators included in the average is not the same for all sites.

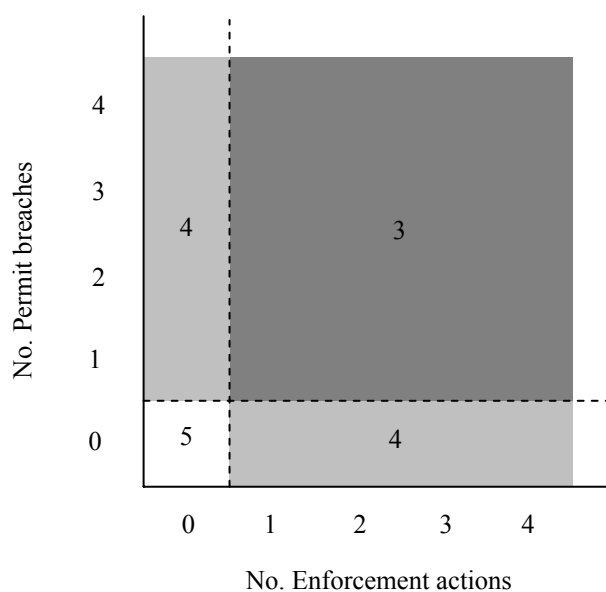
⁴ The maximum number of indicators depends on the particular measure, and on the sector to which the site belongs. For example, for *process efficiency – energy* (EP5), sites in the cement sector were asked to provide data for up to three indicators, while sites in the combustion sector were asked to provide data for up to two.



For *compliance* (EP1) the five indicators represent different measures of (non) compliance – e.g. the number of formal enforcement actions, the number of permit condition breaches, etc. Each indicator is scored on a two-point scale: zero if the site is fully compliant (i.e. it has received no enforcement actions, etc), and five if it is not (i.e. it has received one or more enforcement actions, etc.). The simple average of these scores then represents the number of indicators for which the site is non-compliant. This raw score is then scaled to fit the common index scale (1 – 2).

There are two important points to note about this scoring algorithm. First, it takes no account of the number of instances of non-compliance for each indicator. So, for example, a site with one permit condition breach and one administrative fine will receive a lower score than a site with ten incidents that resulted in significant environmental harm. Second, it implies that there is an equivalence (or trade-off) between the different indicators. This is illustrated in Figure 2.3, which shows the possible scores for different combinations of numbers of enforcement actions and numbers permit condition breaches – assuming that the site is fully compliant for all of the other three indicators.

Figure 2.3: Equivalence of indicators for EP1



For *conduct* (EP2) there are two indicators, representing the number of substantiated complaints received, and the number of unresolved complaints. These are both scored on a three-point scale (0, 2, 4). Consequently, the overall score (which is equal to the simple average of the two scores) is measured on a five-point scale (0, 1, 2, 3, 4). Again, this is scaled to fit the common index scale.

For the process efficiency and releases measures (EP3 – EP8), the indicators are sector-specific. For example, the raw material efficiency indicator for the combustion sector is fuel usage per MWhr power output, while for the cement sector it is use of natural raw materials per tonne of clinker. In each case however, the sites provide actual values for their selected indicators. For each indicator, the actual values are compared with four “cut-off” values, to score the site on a five-point scale (0 – 4); with higher scores



corresponding to higher efficiency, or lower releases.⁵ Each site's overall score is then calculated as the simple average of the scores for its selected indicators, with the resultant value being scaled to the common index scale.

Unlike the algorithm used calculating for *compliance* (EP1) scores, the use of a five-point scale means that some account is taken of differences in the actual values for each indicator. However, the averaging of the scores again implies that there is an equivalence between the different indicators for each measure. This is illustrated in Figure 2.2 for the case of *releases to air* (EP6), where the two axes represent the emission rates (in mg/Nm³) of two different pollutants.⁶ For each pollutant, the dashed lines represent the respective benchmark values, while the dotted lines represent the cut-off values used to determine the individual scores. The cut-off values implicitly define equivalence contours between the two pollutants (shown as the solid lines). The shapes of these contours depend on the choice of the respective cut-off values. If the gap between the cut-off values gets successively larger for each pollutant, then the contours are relatively regular and smooth (i.e. case (a) in Figure 2.4). If this is not so, then the contours are irregular and jagged (i.e. case b), although they will never cross.

Figure 2.4: Equivalence contours for emissions of air pollutants (mg/Nm³)

case (a): regular contours

case (b): irregular contours

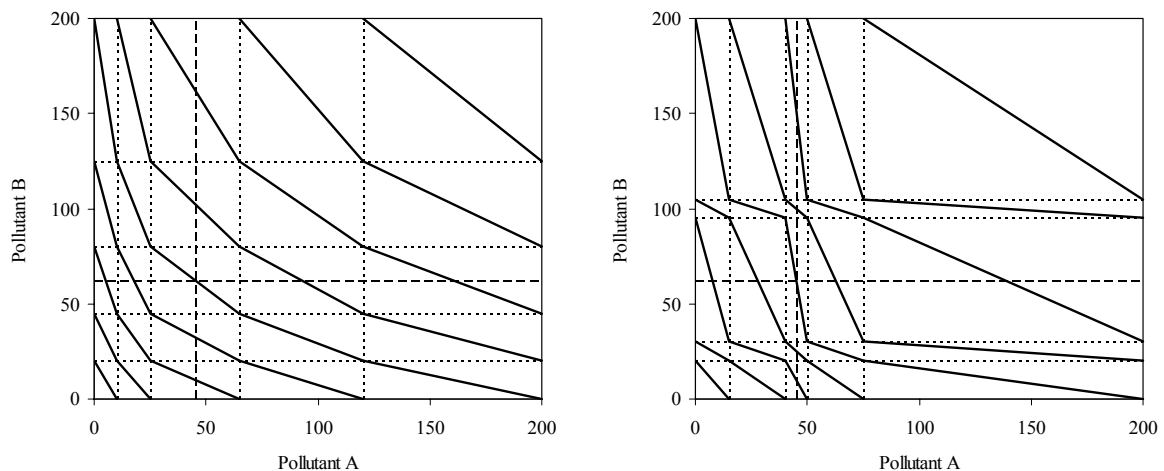


Figure 2.5 shows the distribution of scores for each of the environmental outcome measures. The potential number of different scores for each measure varies between five (for EP2 and EP3) and seventeen (for EP5 – EP8).⁷ However, as can be seen the actual number of different scores observed for the sample sites is typically much lower.

⁵ The cut-off values are determined by reference to the relevant IPPC benchmark value (or values).

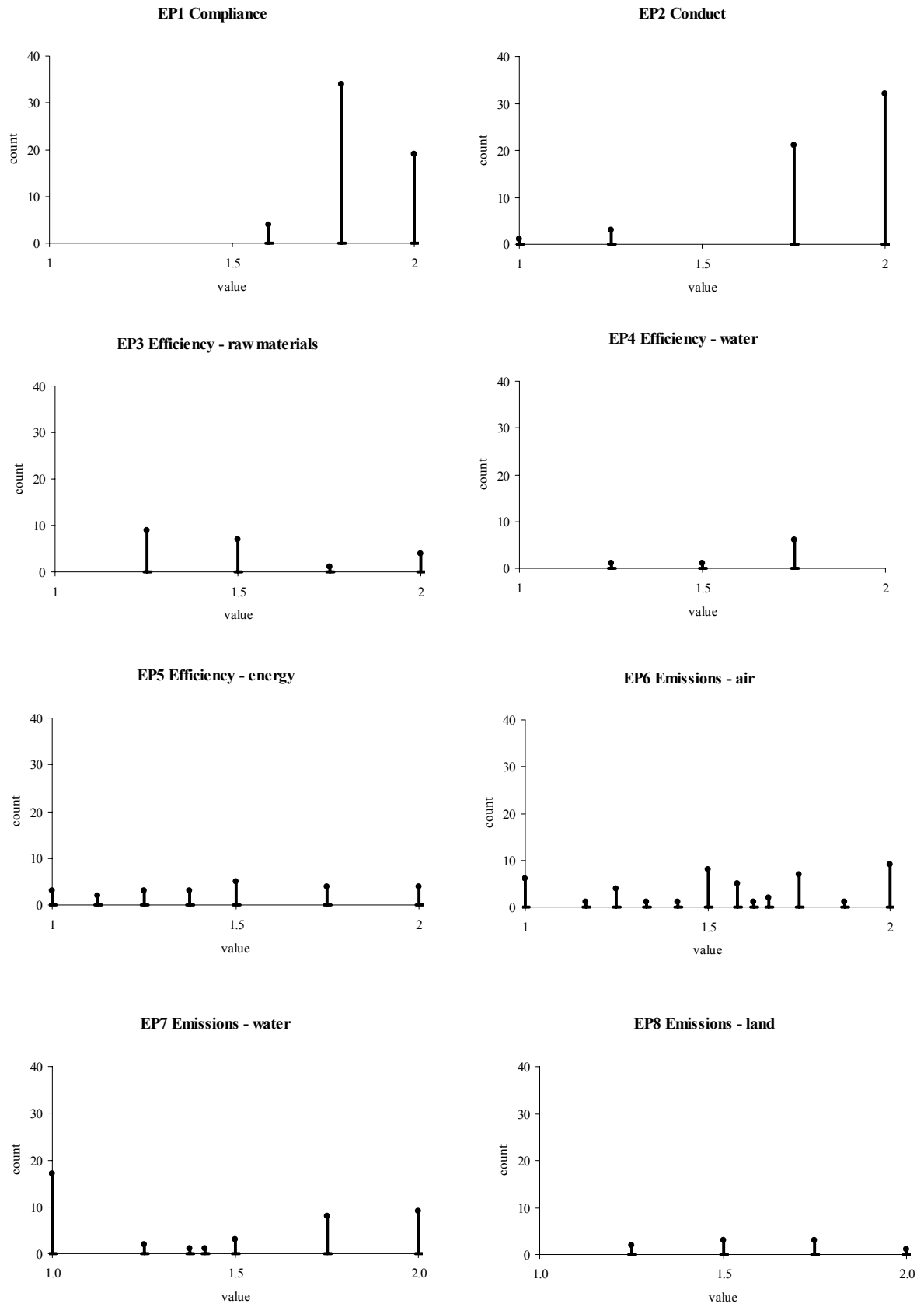
⁶ In this hypothetical example, it is assumed that the maximum possible level of emissions for each pollutant is 200 mg/Nm³.

⁷ The potential number of scores for each measure is equal to $n \times (p-1) + 1$, where n is the number of indicators included in the average, and p is the number of points on the indicator scoring scale. For example, for EP1 there are 5 indicators and 2 points on the scoring scale, and hence the potential number of scores is $5 \times (2-1) + 1 = 6$.





Figure 2.5: Distribution of scores for EP measures



Summary statistics for the eight environmental outcome measures are provided in Table 2.7. Apart from *compliance* (EP1) and *conduct* (EP2), scores have not been provided for all of the sample sites. In particular, only eight of the sample sites have scores for *process efficiency – water* (EP4), and only nine for *releases to land* (EP8). Breakdowns of the sub-samples for EP3-EP8 by industrial sector and EMS classification are given in Appendix C.

Compared to the EMA scores, the EP scores show greater variation – both between and within measures. *Compliance* (EP1) has the highest average score, and the lowest standard deviation; while *releases to water* (EP7) has the lowest average score and the highest standard deviation.

Table 2.7: Summary statistics for environmental outcome measures

Measure	No. of sites	Mean	Minimum	Maximum	Standard deviation	Spread
EP1	57	1.85	1.60	2.00	0.12	0.40
EP2	57	1.85	1.00	2.00	0.22	1.00
EP3	21	1.50	1.25	2.00	0.29	0.75
EP4	8	1.66	1.25	1.75	0.19	0.50
EP5	24	1.48	1.00	2.00	0.33	1.00
EP6	46	1.56	1.00	2.00	0.32	1.00
EP7	41	1.43	1.00	2.00	0.42	1.00
EP8	9	1.58	1.25	2.00	0.25	0.75

For the six measures relating to process efficiency (EP3 – EP5) and releases (EP6 – EP8), the scores are relatively evenly distributed across the range of potential scores. However, the distributions for *compliance* (EP1) and *conduct* (EP2) are both skewed towards the high end of the range, with over 90% of the sample sites achieving one of the top two scores.

Table 2.8 shows the degree of correlation between the scores for the eight environmental outcome measures. It is clear from this that there is no systematic correlation between the measures. The magnitude of the correlation coefficient is greater than 0.6 for only two pairs of measures: *conduct* (EP2) and *process efficiency – water* (EP4); and *releases to air* (EP6) and *process efficiency – water* (EP4); and in both cases the scores are negatively related. Similarly, *compliance* (EP1) is negatively correlated with all three of the measures for releases.

However, caution should be exercised in reading too much into the values of these correlation coefficients. First, a number of the coefficients are based on small sub-samples of the overall sample. In particular, the coefficients highlighted in grey are all based on sample sizes of less than twenty. These include the six coefficients with the largest negative values. Second, if one assumes that the only reason for two EP measures to be positively correlated is that both are positively correlated with operator performance (or some other explanatory variable), then one would not necessarily expect to find a very



strong correlation between the two. For example, for a sample size of fifty, if the expected correlations between operator performance and two particular EP measures are 0.5 and 0.3 respectively, then the expected correlation between the EP two measures is only 0.15 (i.e. 0.5×0.3). For a particular sample, the value of the correlation coefficient may differ considerably from this figure, and it is quite possible for the value to be negative.

Table 2.8: Correlation matrix for EP measures

	EP1	EP2	EP3	EP4	EP5	EP6	EP7	EP8
EP1	1.00							
EP2	0.31	1.00						
EP3	0.00	0.15	1.00					
EP4	0.42	-0.69	0.00	1.00				
EP5	-0.19	0.04	0.18	-0.39	1.00			
EP6	-0.26	0.13	0.11	-0.64	0.04	1.00		
EP7	-0.10	-0.35	-0.48	-0.37	0.03	0.05	1.00	
EP8	-0.28	0.00	0.37	-0.51	0.46	0.01	0.21	1.00

While it is true that if the correlations between operator performance and two EP measures were both very strong (say 0.9), then it is highly likely that one would observe a strong positive correlation between the two measures, the lack of a strong correlation does not necessarily imply operator performance has no impact on either (or both) of the two measures. In order to answer this question, it is necessary to assess the impact for each measure individually. Unfortunately, while the sizes of the sub-samples are sufficiently large to undertake the analysis for six of the eight outcome measures, they are too small to do so for *process efficiency – water* (EP4) and *releases to land* (EP8).

2.3 Impact of EMS classification on operator performance

Table 2.9 summarizes the results of the first stage of the analysis, which considers the relationship between EMS classification and operator performance. More detailed results are given in Appendix D, which provides a detailed description of the analysis.

The first three columns of values in the table are the constant term and the coefficients of the explanatory variables in the calculated *sample regression lines* for overall operator performance (EMA), and for the nine individual dimensions of performance (EMA11-EMA19). These are the lines that provide the “best fit” to the sample data for each measure. The last two columns give the number of sites (n) that are used in the calculation of the coefficients, and the *multiple coefficient of determination* (R^2 value) for the regression. This last measure gives the proportion of the sample variation in the value of the dependent variable (i.e. the operator performance score) that is explained by the sample



regression line. As such it provides an indication of how well the calculated line fits the data. The higher the R^2 value, the better the fit.⁸ With cross-sectional data, low R^2 values are relatively common. However, the value is particularly low for the sample regression line for commitment to training and awareness (EMA13).

Table 2.9: Estimated impact of EMS classification on average operator performance, and on individual dimensions of operator performance

Dep. Variable	Estimated regression coefficients			n	R ²
	Constant	EMAS	No EMS		
EMA	1.70	0.22	-0.26	57	0.41
EMA11	1.77	0.17	-0.21	57	0.22
EMA12	1.77	0.00	-0.33	57	0.21
EMA13	1.61	0.15	-0.01	57	0.03
EMA14	1.78	0.15	-0.41	57	0.29
EMA15	1.67	0.30	-0.17	57	0.40
EMA16	1.54	0.45	-0.30	57	0.41
EMA17	1.75	0.11	-0.39	57	0.27
EMA18	1.69	0.33	-0.35	57	0.27
EMA19	1.58	0.46	-0.29	57	0.40

The constant term represents the expected operator performance score of a site with ISO 14001. The coefficient for “EMAS” represents the difference between the expected score of a site with EMAS and the expected score of a site with ISO 14001. Similarly, the coefficient for “No EMS” represents the difference between the expected score of a site without a recognised EMS and the expected score of a site with ISO 14001. Since the estimation process takes into account the influence of operator performance on EMS classification (see Figure 2.2), the two coefficients measure the causal impacts of the two types of EMS for the sample sites.

Thus, for the sites in this sample, the adoption of ISO 14001 causes the expected value of the overall performance score (EMA) to increase by 0.26, while the adoption of EMAS causes it to increase by 0.48 (i.e. 0.26 + 0.22). A similar picture can be seen for the individual dimensions of operator performance, although the magnitudes of the impacts vary between dimensions. The impact of ISO 14001 ranges from +0.01 for *commitment to training and awareness* (EMA13) to +0.41 for *compliance and conformance control* (EMA14). The incremental impact of EMAS ranges from zero for *operational and risk management* (EMA12) to +0.46 for *environmental reporting* (EMA19).

⁸ The basic concepts of regression analysis, and the various summary measures of “goodness-of-fit” for the sample regression line are explained in Appendix F.



It is important to recognize that the calculated coefficients and constant terms given in Table 2.9 are specific to this sample. If another sample were to be used, then different values would be obtained. Thus, the calculated values only provide an estimate of the true values for the population as a whole (i.e. for all IPPC sites); and, like all estimates, are subject to error.⁹ However, it is possible to use the sample information to test hypotheses about the population values. In this case the relevant hypotheses are that the population value of “EMAS” coefficient in each equation is greater than zero, and that the value of the “No EMS” coefficient is less than zero.¹⁰ For each coefficient, the hypothesis is tested by comparing a *test-statistic* with a *critical value*. The test statistic is equal to the sample value of the coefficient divided by its standard error. The critical value depends on the sample size, the number of variables included in the equation, and *level of significance* that is chosen for the test. This represents the maximum acceptable value for the probability of accepting the hypothesis when it is in fact false. If the test-statistic is greater than the critical value, then the hypothesis is accepted, and the sample value of the coefficient is said to be *significant* at the chosen level of significance.¹¹

A 10% level of significance has been throughout the analysis, and those coefficients that are significant at this level are highlighted in grey in Table 2.8 (similarly in Table 2.10 and Table 2.11). The coefficients in the sample regression line for overall operator performance (EMA) are both significant. Indeed, as can be seen from Table D4 in Appendix D, they are actually significant at a 2% level of significance. Thus, the sample provides strong evidence to support the hypothesis that the adoption of an accredited environmental management system leads to an overall improvement in operator performance (as measured by a site’s average EMA score); with EMAS having a greater beneficial impact than ISO 14001.

In the regression lines for the individual dimensions of operator performance, all but four of the coefficients are highly significant, and two of these (the EMAS coefficients for EMA13 and EMA17) only just fail the test. Thus the hypothesis is supported for seven of the nine dimensions of operator performance. However, the sample provides no evidence that the adoption of ISO 14001 leads to an improvement in a site’s *commitment to training and awareness* (EMA13), or that the adoption of EMAS leads to an incremental improvement over ISO 14001 for *operational and risk management* (EMA12). Of course, this does not necessarily mean that the actual population values of these coefficients are not positive; just that this particular sample does not provide any evidence that this is the case.

2.4 Impact of operator performance on environmental outcomes

The relationship between operator performance and the various environmental outcomes is considered in the second stage of the analysis. Apart from *compliance* (EP1) and *conduct* (EP2), the algorithms that have been used to generate the EP scores are sector-specific. It

⁹ Provided that certain conditions are met, the expected value of this error is equal to zero. That is, if the coefficient values were to be calculated for a large number of different samples, then the averages of the values across all of the samples would be equal to the population values.

¹⁰ For hypotheses of this type, a “one-tail” test is used. If the hypothesis had been that the value of the coefficient is not equal to zero (i.e. no assumption is made about its sign, as is the case with the coefficients of the various site characteristic variables, then a “two-tail” test is used.

¹¹ A more detailed explanation of hypothesis testing and levels of significance is given in Appendix F.



seems plausible therefore that the relationships between operator performance and process efficiency, and between operator performance and releases, may vary from sector to sector. Unfortunately, with a sample size of fifty-seven, it is not possible to allow for potential differences in the impacts of all nine individual dimensions of operator performance.¹² Consequently, the analysis is broken down into two parts. In the first part the impact of overall operator performance (as measured by the average EMA score) on each environmental outcome is assessed, allowing for differences between the sectors. In the second part, all nine dimensions of operator performance are included in the regression function, but no allowance is made for differences between sectors.

The results of the first part of the analysis are given in Table 2.10, which shows the sample values of the regression coefficients for overall operator performance (EMA), for each of the six environmental outcome measures for which there is sufficient data. The values of the constant terms and the other coefficients in the respective equations are given in Tables E1-E6 in Appendix E, which provides a detailed description of the stage-two analysis. As was the case in Table 2.9, coefficients that are statistically significant (at a 10% level of significance) are highlighted in grey. While the values are reported separately for each sector, they are calculated using a single “pooled” regression.¹³ Where the analysis finds no significant differences between two (or more) sectors, they are amalgamated and a common value is calculated for the EMAS coefficient.

Table 2.10: Estimated impact of average operator performance on environmental outcomes (sample sizes in brackets)

Measure	Cement	Combustion	Organic chemicals	Inorganic chemicals	Paper & pulp	Other
EP1	-0.13 (57)	-0.13 (57)	-0.83 (57)	-0.13 (57)	-0.13 (57)	-0.13 (57)
EP2	-0.74 (18)	-0.39 (39)	-0.39 (39)	-0.39 (39)	0.55 (18)	-0.39 (39)
EP3	0.62 (13)	- -	- -	- -	0.73 (8)	0.73 (8)
EP5	2.84 (19)	-0.05 (19)	- -	- -	1.94 (5)	-0.05 (19)
EP6	2.34 (46)	-0.13 (46)	-0.13 (46)	-0.13 (46)	-0.13 (46)	-0.13 (46)
EP7	-1.27 (41)	1.54 (41)	1.54 (41)	-1.33 (41)	1.54 (41)	1.54 (41)

¹² To do so for six sectors would require the inclusion of seventy-two variables in the regression line (including the constant term), which is greater than the number of observations for the sample.

¹³ For technical reasons, sectors are grouped into two pools for EP2, EP3 and EP5; with separate equations being estimated for each pool.



It is clear from Table 2.10 that the sample provides no evidence to support the hypothesis that higher levels of operator performance lead to better *compliance* (EP1), or to better *conduct* (EP2). None of the sample values of the EMA coefficients for these measures are significant, and all but one are negative.¹⁴ However, this may reflect the construction of the EP scores for these measures; particularly in the case of EP1, which takes no account of the degree of non-compliance. Thus, all else being equal, a site with ten formal enforcement actions against it in the last twelve months is given the same score as a site with only one. This issue is discussed more fully in section 3.

For the other four measures, the picture is mixed. For the two process efficiency measures (EP3 and EP5), five of the seven values are positive, although only one is significant at a 10% level of significance. However, it should be noted that because of the need to split the sectors into two pools, the sample sizes on which these calculations are based (shown in brackets) are very small. For the two releases measures (EP6 and EP7), five out of twelve values are positive. However, all of these are significant. Thus, while the sample provides some support for the hypothesis that higher levels of operator performance leader to better environmental outcomes in terms of process efficiency and releases, the evidence is not very strong.

The results of the second part of the analysis are given in Table 2.11. Again, the coefficients are only shown for the nine dimensions of operator performance. The values of the constant terms and the other coefficients in the respective equations can be found in Tables E7-E12 in Appendix E.

Table 2.11: Estimated impact of individual dimensions of operator performance on environmental outcomes

	Regulation			Process efficiency			Releases		
	EP1	EP2	Ave	EP3	EP5	Ave	EP6	EP7	Ave
EMA11	0.08	-0.13	-0.03	1.82	2.70	2.26	-0.87	-0.47	-0.67
EMA12	-0.04	0.44	0.20	-0.32	-0.22	-0.27	0.47	-1.26	-0.40
EMA13	-0.06	-0.07	-0.07	-0.34	0.66	0.16	0.08	0.07	0.08
EMA14	0.14	-0.32	-0.09	0.80	0.56	0.68	-0.28	0.97	0.35
EMA15	0.00	-0.40	-0.20	-1.00	-1.29	-1.15	0.24	1.52	0.88
EMA16	-0.10	-0.11	-0.11	-0.85	0.35	-0.25	-0.95	0.07	-0.44
EMA17	-0.07	-0.14	-0.11	-0.65	-0.20	-0.43	-0.44	0.12	-0.16
EMA18	-0.07	0.25	0.09	0.31	-1.93	-0.81	0.56	-0.34	0.11
EMA19	-0.05	-0.24	-0.15	-0.34	0.42	0.04	0.75	0.37	0.56
n	57	57		21	24		46	41	
R ²	0.13	0.38		0.74	0.60		0.21	0.44	

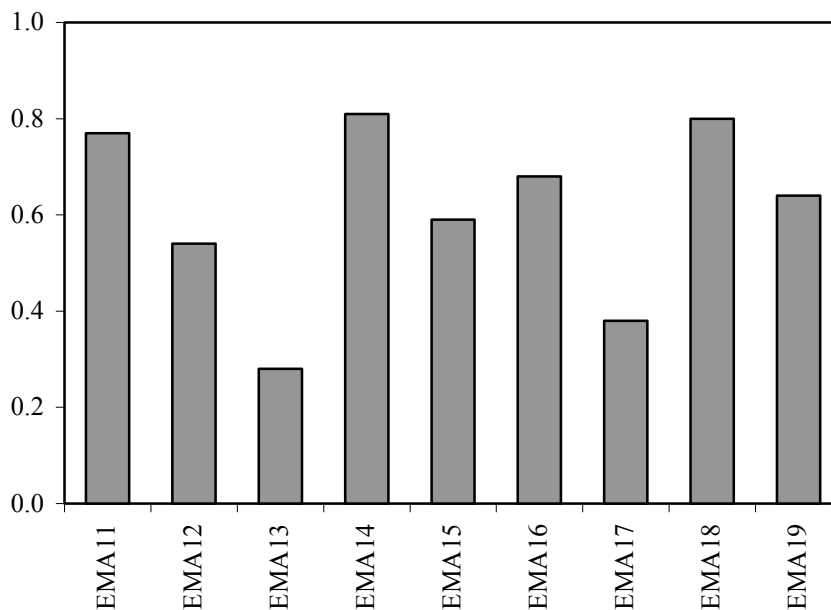
¹⁴ As has been noted previously, a negative sample value for a coefficient does not necessarily imply that the corresponding population value is negative.



Only six of the fifty-four coefficient values are significant at a 10% level of significance, and more than half are negative. However, this is not entirely unexpected given the high degree of inter-correlation between the independent variables – a problem known as *multicollinearity*. All else being equal, if a variable is strongly correlated with the other independent variables in the regression equation, then its sample coefficient value will have a large standard error; reducing the value of the test statistic, and making it less likely that the hypothesis (of a positive coefficient value) will be accepted when it is in fact true.

The strength of the correlation for each variable is given by the R^2 value for the “auxiliary regression” of that variable on all of the other independent variables in the main regression equation.¹⁵ These values are shown in Figure 2.6, from which it is clear that the degree of the problem varies between the different dimensions of operator performance. While the R^2 values are particularly high for *environmental policy* (EMA11), *compliance and conformance control* (EMA14) and *management review* (EMA18), multicollinearity is not a significant problem for *commitment to training and awareness* (EMA13), or for *documentation control* (EMA17).¹⁶

Figure 2.6: Auxiliary R^2 values for operator performance measures



Nevertheless, it is possible to discern some patterns in the coefficient values; particularly if one allows for differences between the three different types of outcome.

- With the possible exception of *operational and risk management* (EMA12), none of the dimensions of operator performance appear to have any impact on the outcome in

¹⁵ For example, in the auxiliary regression for EMA11, the independent variables are EMA12-EMA19 and the two size-related variables

¹⁶ The R^2 values shown in Figure 2.6 are consistent with the simple “pair-wise” correlation coefficients between the different dimensions of operator performance that are shown in Table 2.5. However, it should be noted that, in general, it is possible to have significant multicollinearity even though all of the simple correlation coefficients are relatively low.



terms of *regulation* (EP1 and EP2). All of the other coefficients are insignificant, and the majority are negative. This is, of course, not too surprising given the lack of any discernable impact of overall operator performance (EMA) on these measures.

- For *process efficiency* (EP3 and EP5), *environmental policy* (EMA11) has the greatest impact, followed by *compliance and conformance control* (EMA14). The sample values of the coefficients for EMA11 are significant for both outcome measures; and while neither of the coefficient values are significant for EMA14, they are both positive with a similar magnitude.
- For *releases* (EP6 and EP7), *performance monitoring* (EMA15) appears to have the greatest impact, followed by *environmental reporting* (EMA19) and *compliance and conformance control* (EMA14). For EMA15 and EMA19, the coefficient values are positive for both outcome measures, and significant for one of measures. For EMA14, the coefficient value is significant for one of the measures, but negative for the other.

Interestingly, *environmental policy* (EMA11) and *compliance and conformance control* (EMA14) are two of the variables that suffer most from the problem of multicollinearity.

None of the other four dimensions of operator performance appear to have any impact. In the case of *open communication culture* (EMA16) and *management review* (EMA18), this may be due to high levels of multicollinearity. However, this is not a significant problem for *training and awareness* (EMA13), or for *documentation control* (EMA17). Consequently, one can reasonably conclude that the sample does not provide any evidence to support the hypothesis that either of these two dimensions of operator performance has any impact on any of the environmental outcomes.

These observations seem intuitively plausible. In particular, it does not seem unreasonable that the relative impacts of the different dimensions of operator performance should vary between the different types of environmental outcome. However, they should be treated with a considerable degree of caution. When there is a high level of multicollinearity, the values of the parameter estimates (and their standard errors) can be very sensitive to small changes in the sample data. It is possible therefore that the addition of some new sample observations may cause some of the significant coefficient values to become insignificant, and / or the positive values to become negative (and vice versa in both cases).



3. Conclusion

In this analysis, information provided for the fifty-seven sample sites has been used to test hypotheses about the underlying relationships between EMS classification, operator performance and environmental outcomes for the population as a whole (i.e. for all industrial sites subject to IPPC). In particular, that the adoption of a formal (accredited) environmental management system leads to higher levels of operator performance, and that higher levels of operator performance lead to better environmental outcomes.

The sample provides strong evidence to support the hypothesis that the adoption of an accredited environmental management system leads to an overall improvement in operator performance (as measured by a site's EMA scores); with EMAS having a greater beneficial impact than ISO 14001. However, it provides no evidence to support the hypothesis that higher levels of operator performance lead to better environmental outcomes in terms of *compliance* (EP1) and *conduct* (EP2), and only very weak evidence that these lead to better outcomes in terms of process efficiency (EP3-EP5) and releases (EP6-EP8).

Essentially there are three possible explanations for the apparent lack of a positive relationship between operator performance and environmental outcomes

- there is no relationship between the two in reality;
- there is a relationship in reality, but we have not been able to detect it with this sample;
- the EP measures used in the analysis do not provide a good representation of the actual environmental outcomes.

It is of course possible that there is no relationship between operator performance and some, or all, of the environmental outcomes. Indeed, a previous study undertaken by PSI for the Environment Agency failed to find any relationship between operator performance and regulatory compliance.¹⁷ However, another study of industrial sites in Mexico did find a positive relationship between the presence of ISO14001 and (self-reported) regulatory compliance.¹⁸ Therefore it seems a little premature to accept this as the explanation. Furthermore, given the objective of the study, this possibility should only be considered as a last resort, when (and if) the other two explanations have been investigated and eliminated. These are considered in turn.

3.1 Incorrect inferences from sample values

In this analysis, the values of the various regression coefficients that have been calculated from the sample data have been used to test hypotheses about the signs of the corresponding population values. In most cases, the hypothesis that has been tested is that

¹⁷ "Environmental Management Systems and Operator Performance at Sites under Integrated Pollution Control", K. Dahlström & J. Skea, Policy Studies Institute, November 2002.

¹⁸ "What Improves Environmental Performance? Evidence from Mexican Industry", S. Dasgupta, H. Hettige & D. Wheeler, World Bank Development Research Group, Working Paper Series #1877, December, 1997



the population value of the coefficient is greater than zero; indicating that the respective explanatory variable has a positive impact on the dependent variable in question.¹⁹

When considering the implications of the analysis it is important to understand the intrinsic nature and limitations of the hypothesis testing procedure. This is important for two reasons. First, it provides some context for the results, thus reducing the possibility of incorrect conclusions being drawn. Second, it enables an assessment to be made of the potential impacts on the results of a number of actions that might be taken.

Ideally, one would want the test procedure to accept the hypothesis whenever it is true (i.e. when the population value of the coefficient really is greater than zero), and to reject it whenever it is false. That is, the test procedure should always lead to the correct conclusion being drawn. Unfortunately, it is technically impossible to achieve this ideal outcome, and hence for any test procedure there is always a chance that conclusion will be incorrect. In particular, there are two types of error that can occur:

- the test procedure may conclude that the population value of the coefficient is greater than zero, when in reality it is not;
- the test procedure may conclude that the population value of the coefficient is not greater than zero, when in reality it is.

Thus, one cannot draw a definitive conclusion about the truth, or otherwise, of the hypothesis from the test result. The most that one can say is that that the hypothesis is likely to be true, or false, on the basis of the sample information. In particular, it is perfectly possible for the actual population value of a particular regression coefficient to be positive when the calculated sample coefficient is negative (as is the case with a number of the coefficients shown in Table 2.10 and Table 2.11).

The first error is known as a “Type I error”.²⁰ The probability of this error occurring is called the level of significance for the test, and it is set by the analyst. The value that is chosen for a particular analysis will depend on the consequences of incorrectly concluding that the coefficient value is greater than zero (i.e. that the explanatory variable has a positive impact). If these are serious, then a low value should be set; if they are not, then a higher value may be acceptable. Typically, the significance levels that are used range from 0.01 to 0.1 (i.e. from 1% to 10%) – although there is nothing sacrosanct about these values. As has already been noted in section 2, a 10% significance level has been used for all the hypothesis tests conducted in this analysis.

The second error is known as a “Type II error”. The probability of not committing this error is called the power of the test. Unlike the level of significance, the power of the test is

¹⁹ In order to simplify the discussion it will be assumed that the hypothesis is always that the value of the population coefficient is greater than zero. However, this does not have to be the case. For some variables (such as the indicator variable for No EMS in the first part of the analysis) the appropriate hypothesis is that the coefficient value is negative. When there are no prior expectations about the impact of a particular variable (such as the site characteristics), the appropriate hypothesis is that the value is not equal to zero.

²⁰ Technically, the hypothesis that is tested is that the population coefficient is less than or equal to zero. This is called the “null hypothesis”, and it represents the opposite of the hypothesis that we are interested in validating. Thus, a Type I error occurs if the test procedure rejects the null hypothesis when it is, in fact, true. Similarly, a Type II error occurs if the test procedure fails to reject the null hypothesis when it is, in fact, false.

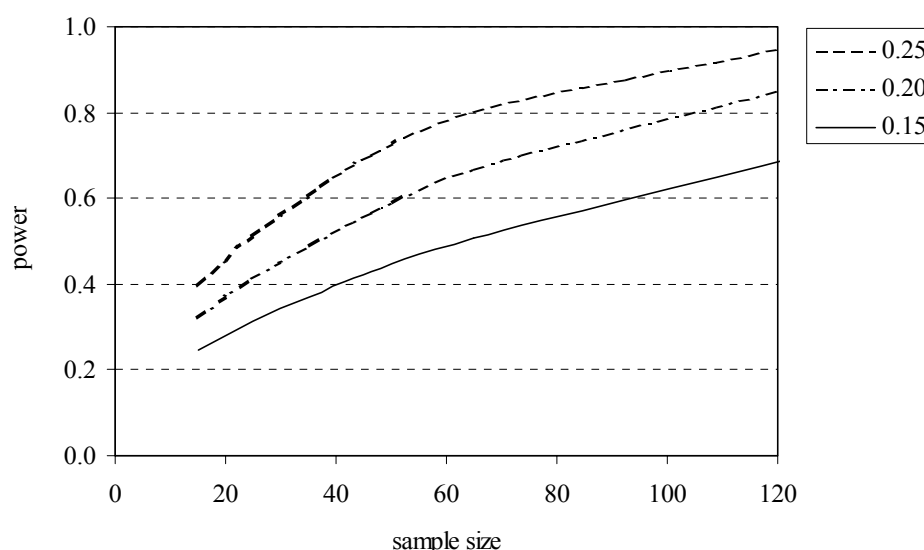


not controlled by the analyst. It depends on the actual value of population coefficient, the level of significance that has been set for the test, and the precision of the estimated coefficient value. The first of these factors is of course fixed, and in practice so too is the second. Consequently, the only way to improve the power of the test is to improve the precision of the estimated coefficient value. Essentially there are two things that can be done to achieve this:

- increase the size of the sample;
- increase the degree of variation in the sample values of the explanatory variable.

Figure 3.1 shows the impact of these two actions on the power of the hypothesis test for a representative example. While this is based on a hypothetical regression equation, the parameter values are typical of those applying to the various equations estimated from the sample data.²¹

Figure 3.1: Impact of sample size and standard deviation of sample values for explanatory variable on power of hypothesis test



The sample coefficients reported in Table 2.10 and 2.11 are based on sample sizes that range from 5 to 57 (depending on the particular EP measure, and the degree of sectoral segmentation), while the standard deviations of the EMA scores vary between 0.15 (for EMA and EMA15) and 0.25 (for EMA18 and EMA19). The powers of the respective hypothesis tests therefore range between 20% and 70% – with an average of around 50%. Thus, where the tests have concluded that the population value of the coefficient is not greater than zero, there is probably around a one in two chance that this conclusion is incorrect.

²¹ The regression equation has five variables (including a constant), and the standard deviation of the population disturbance term is equal to 0.25. The actual population value of the coefficient under consideration is 0.5, and the R^2 value of the auxiliary regression for the variable is 0.7. The level of significance for the hypothesis test is set at 0.1.



The acceptable minimum value for the power of the hypothesis tests is a matter of judgement. However, given that the level of significance for the tests has been set at 0.1, a target value of around 0.9 seems reasonable. At this level, the chances of the two types of error occurring are equalized, at one in ten. As can be seen in Figure 3.1, this can be achieved by increasing the sample sizes so that they range between 100 and 120, and increasing the variability of the EMA scores so that the standard deviations lie in the range 0.2 to 0.25.

The total sample size required to achieve the target size range will depend on the sample penetration rates that are achieved for each measure. The penetration rates for the current sample are shown in Table 3.1. As can be seen, there is considerable variation between the different measures – with 80% of the sample sites having scores for *releases to air* (EP6), but only 14% having scores for *process efficiency – water* (EP4). To a large extent this variation reflects the fact that IPPC benchmark values are not available for all measure / sector combinations. Consequently, for those combinations without benchmark values, it has not been possible to calculate scores even though sites may have reported the raw data.

Table 3.1: Proportion of sample sites with EP scores , by sector

Industrial sector	Process efficiency			Releases			Total
	EP3	EP4	EP5	EP6	EP7	EP8	
Cement	100%	-	100%	100%	85%	-	13
Combustion	-	-	11%	100%	67%	-	9
Inorganic chemicals	-	-	-	83%	83%	-	6
Organic chemicals	-	-	-	58%	75%	-	12
Paper & pulp	80%	100%	100%	80%	100%	80%	5
Other	33%	25%	42%	67%	42%	42%	12
Total	37%	14%	42%	80%	72%	16%	57

If the minimum penetration rate could be increased to 50%, then a total sample size of 200 to 240 would be sufficient to meet the target size ranges. In order to achieve this, it may be necessary to use alternative reference values for those combinations of measures and sectors for which there are no IPPC benchmark values. For example, the reported values could be compared with the average for the sample, or with an industry average value, in order to assign scores.

Ideally, there should be roughly equal numbers of sites in each industrial sector, as it may be necessary to split the sample into two (or more) “sector pools” for some of the analyses. However, as the total sample size increases, this becomes less critical.

Given the strong relationship between EMS classification and operator performance (see Table 2.9), the variability of the EMA scores will be increased if the proportions of the sample sites with EMAS, and with no EMS are increased. As can be seen in Table 3.2, these classifications account for 16% and 30% of the current sample. If they could be increased to 30% and 40% respectively, then the target increase in the standard deviation



would be achieved. For example, this shift in the sample mix would increase the standard deviation of the *overall operator performance* score (EMA) from 0.15 to 0.19.

Table 3.2: Number of sample sites by sector / EMS classification, by sector

Industrial sector	EMS classification			Total %	Total number
	EMAS	ISO 14001	No EMS		
Cement	62%	23%	15%	100%	13
Combustion	0%	89%	11%	100%	9
Inorganic chemicals	0%	67%	33%	100%	6
Organic chemicals	8%	42%	50%	100%	12
Paper & pulp	0%	60%	40%	100%	5
Other	0%	67%	33%	100%	12
Total	16%	54%	30%	100%	57

Therefore when increasing the sample size, the objective is to create a more balanced sample, with greater variation in operator performance scores. However, it is important that sites are not selected on the basis of environmental outcomes. To avoid this possibility, the EP scores should only be calculated for sites once they have been selected for the sample

3.2 Distorted representation of environmental outcomes

An important assumption underpinning the analysis is that the EP measures provide a good representation of the respective environmental outcomes. Clearly, if this is not the case, then the results of the analysis may be adversely affected. There would appear to be three potential issues that could affect the validity of this assumption:

- the choice of methodology for combining indicators;
- the treatment of “missing” indicators;
- the choice of definitions for indicators.

The first issue is probably the most fundamental of the three. As has been described in section 2.2, where measures are based on more than one indicator, the overall scores are calculated as the simple averages of the respective indicator scores. Underlying this methodology is an implicit assumption that there is an equivalence (or trade-off) between the component indicators. That is, a poorer outcome for one of the indicators can be offset by a better outcome for one (or more) of the others.

This average scoring methodology is appropriate when the component indicators are comparable, and can be measured in common units – or can be converted into a common unit. As such it seems reasonable to use it for the calculation of the process efficiency measures and the emissions measures. However, for the two regulation measures, where



the components are not directly comparable, it is not clear that this methodology is the most appropriate. In particular, for *compliance* (EP1) an alternative methodology – based on the principle of lexicographic ordering – may give a better reflection of the relative outcomes. Under this methodology, the component indicators are ranked in order of importance. Sites that are not fully compliant for the most important indicator are given the lowest possible score; sites that are non-compliant on the second most important indicator are given the next lowest score; and so on. The difference between the two methodologies is illustrated in Table 3.3, for the case of three indicators (and hence a maximum possible score of 3).

Table 3.3: Average and lexicographic scoring methodologies

Overall score	Average scoring	Lexicographic scoring
3	A: X1 = 0 X2 = 0 X3 = 0	A: X1 = 0 X2 = 0 X3 = 0
2	B: X1 = 0 X2 = 0 X3 > 0	B: X1 = 0 X2 = 0 X3 > 0
	C: X1 = 0 X2 > 0 X3 = 0	
	D: X1 > 0 X2 = 0 X3 = 0	
1	E: X1 = 0 X2 > 0 X3 > 0	C: X1 = 0 X2 > 0 X3 = 0 E: X1 = 0 X2 > 0 X3 > 0
	F: X1 > 0 X2 = 0 X3 > 0	
	G: X1 > 0 X2 > 0 X3 = 0	
0	H: X1 > 0 X2 > 0 X3 > 0	D: X1 > 0 X2 = 0 X3 = 0
		F: X1 > 0 X2 = 0 X3 > 0
		G: X1 > 0 X2 > 0 X3 = 0
		H: X1 > 0 X2 > 0 X3 > 0

The number of possible scores is the same in each case.²² However, there are clearly some important differences between the two methodologies. First, the relative ordering of the outcomes can be different. For example, outcome D is given a higher score than outcome E under the average scoring methodology, but a lower score under the lexicographic methodology. Second, apart from the highest score (i.e. 3), the number of possible outcomes for each score is different. In particular, there are four outcomes that are given the lowest score (i.e. 0) under lexicographic scoring, but only one under average scoring. This is reflected in the distribution of the scores under the two methodologies.

Table 3.4 shows the distribution of EP1 scores under the two methodologies for an illustrative sample of one thousand sites, where the values of the five indicator variables for each site have been generated randomly.²³ As can be seen, the distribution of scores under the average scoring methodology – using a two point scale for the component indicators – is very similar to that observed for the fifty-seven sample sites in this analysis (see Figure 2.3). There are two points to note. First, differentiating between degrees of

²² This reflects the use of a two-point scale for the indicators under the average scoring methodology, and the particular lexicographic ordering rule that has been adopted. It would not be the case, for example, if a five-point scale had been used for the indicators.

²³ For each site, the number of instances of non-compliance for each of its component indicators is drawn from an independent Poisson distribution, with mean equal to 0.2.



non-compliance for each of the component indicators has very little impact on the distribution of the scores. When indicators are scored on a five-point scale, the distribution is only marginally different to that arising under the two-point scale. In contrast, the adoption of the lexicographic methodology has a significant impact on the distribution. While the proportion of sites fully-compliant sites (receiving the highest score) is unchanged, the distribution of the non-compliant sites across the other five possible scores is much more even.

Table 3.4: Distribution of EP1 scores under average and lexicographic scoring methodologies

Overall score	Average scoring		Lexicographic scoring
	Two-point	Five-point	
5	35%	35%	35%
4	42%	38%	9%
3	20%	21%	10%
2	3%	5%	12%
1	0%	1%	18%
0	0%	0%	16%

Neither of the two methodologies is intrinsically better (more appropriate) than the other. Indeed they are not the only methodologies that could be used. As noted above, the averaging methodology implicitly assumes that there is a trade-off between the component indicators. In contrast, the lexicographic methodology assumes that there is no trade-off. Ultimately, the choice between the methodologies it is a matter of judgement as to which better reflects the “concept” of compliance.

The second issue arises in the calculation of the average scores for the process efficiency measures (EP3 – EP5) and the releases measures (EP6 – EP8). Because sites are allowed to choose how many indicator values to provide for these measures, the number of indicators included in the calculation of the average score will vary from site to site. Implicit in this approach is an assumption that if the site had provided a value for another indicator, the score for that indicator would have been equal to the average of the scores for the indicators for which it did provide data. However, this assumption is only valid if there is a strong correlation between the indicator values / scores. If this is not the case, then the approach may give misleading picture of the relative environmental outcomes.

This possibility is illustrated in Table 3.5, which shows the calculation of the average scores for releases to air (EP6) for four sites. In this hypothetical example, there are three potential indicators. Three of the sites have provided values for all three indicators, while site A has only provided values for two. Under the adopted approach, site A has the highest average score. However, this does not seem justified. For the two indicators for which it has provided values, it has higher emissions than site B, and the same emissions as site C. Dividing site A’s total score by three (which is the same as replacing the missing score with 0) would be equally misleading. This would give it the same average score as site D, which has higher emissions for both indicators.



Unfortunately, it is not clear what should be done in this situation. An alternative approach might be to use only the most important indicator for each site. This would avoid the problem highlighted in Table 3.5. However, it would waste potentially valuable information, and there would be the problem of determining which of the indicators is most important.

Table 3.5: Calculation of EP6 scores

	Indicator 1		Indicator 2		Indicator 3		Average score
	Value (mg/Nm ³)	Score	Value (mg/Nm ³)	Score	Value (mg/Nm ³)	Score	
Site A	80	2	n/a		30	3	2.50
Site B	67	3	67	0	20	4	2.33
Site C	80	2	40	1	30	3	2.00
Site D	100	1	20	3	50	1	1.67

The final issue is particularly relevant to the treatment of unresolved complaints in the calculation of the measure for *conduct* (EP2). There are two definitions that could be used for this indicator: (1) the total number of unresolved complaints; or (2) the percentage of substantiated complaints that are unresolved (i.e. the total number of unresolved complaints divided by the total number of substantiated complaints received). As the simple example in Table 3.6 illustrates, the choice between these two alternatives can have a significant impact on the relative ordering of sites.

Table 3.6: Comparison of alternative definitions for unresolved complaints

	Number of complaints	Score	Number unresolved	Score	Percentage unresolved	Score
Site A	300	0	60	0	20%	2
Site B	200	1	50	1	25%	1
Site C	100	2	40	2	40%	0
Average	200		50		25%	

In this example, site A has the greatest number of complaints, and the greatest number of unresolved complaints. However, it has the lowest proportion of unresolved complaints (i.e. it is the best at resolving any complaints that are received). In contrast, site C has the lowest number of complaints and unresolved complaints, but the worst resolution performance. If the number of unresolved complaints is used in the calculation, then site C receives the highest score, and site A the lowest. However, if the percentage of complaints unresolved is used, then the outcome depends on the relative weights that are given to the two component indicators. In particular, if more weight is given to the



percentage of complaints unresolved, then site A receives the highest score, and site B the lowest.

As with the choice between methodologies, the choice of appropriate definition is a matter of judgement. However, depending on which one is used, it is possible that the calculated EP2 score of a particular site may be higher, or lower, than another. Clearly, if there are significant differences between the relative scores of sites under the two definitions, then the choice of definition may affect the results of the analysis.

3.3 Recommendations

In conclusion, the following actions would enhance the analysis of the relationships between the various dimensions of operator performance and environmental outcomes.

1. Increase the sample size for each EP measure to a minimum of around 100 sites – either by increasing the total sample size, or by increasing the sample penetration for each measure, or both.
2. Increase the variability of the EMA scores for each dimension of operator performance – by increasing the proportions of sample sites with EMAS, and with no EMS.
3. Review the methodology that is used to calculate the *compliance* measure (EP1) to ensure that it properly reflects the “concept” of regulatory compliance
4. Review the definition of the indicator for unresolved complaints that is used in the calculation of the *conduct* measure (EP2), and (possibly) the methodology that is used for combining the two indicators
5. Review the correlation between the indicator scores for each of process efficiency measure (EP3 – EP5) and each releases measure (EP6 – EP8), to check the validity of the treatment of “missing” indicator values.
6. Review the cut-off values that are used to assign scores for the indicators used in the calculation of each of process efficiency measure (EP3 – EP5) and each releases measure (EP6 – EP8), to check that the implied “equivalence contours” between indicators are correct.

There is, of course, no guarantee that these actions will result in the detection of a positive relationship between operator performance and environmental outcomes. However, they should enable a more definitive conclusion to be drawn, one way or the other.



4. Appendices

- Appendix A: Allocation of Environmental Management Assessment questions to EMA measures
- Appendix B: Definition of variables used in the analysis
- Appendix C: Breakdown of sample sites with EP scores for process efficiency and releases
- Appendix D: Stage 1 analysis
- Appendix E: Stage 2 analysis
- Appendix F: Interpretation of regression results



Appendix A: Allocation of Environmental Management Assessment questions to EMA measures

The scores for the environmental management (EMA) measures are based on the sites' responses on an Environmental Management Assessment questionnaire. This comprises fifty-six questions, broken down into the following six broad areas:

- EMA1 Environmental policy
- EMA2 Planning
- EMA3 Implementation and operation
- EMA4 Checking and corrective action
- EMA5 Management review
- EMA6 Reporting environmental performance

The questions were subsequently reallocated to the nine dimensions of environmental management used in the analysis. That is:

- EMA11 Environmental policy
- EMA12 Operational and risk management
- EMA13 Commitment to training and awareness
- EMA14 Compliance and conformance control
- EMA15 Performance monitoring
- EMA16 Open communication culture
- EMA17 Documentation control
- EMA18 Management review
- EMA19 Reporting environmental performance

Table A1 provides details of the reallocation.



Table A1: Reallocation of Environmental Management Assessment questions

Questionnaire		Environmental management dimensions used in analysis								
Area	Question	EMA11	EMA12	EMA13	EMA14	EMA15	EMA16	EMA17	EMA18	EMA19
EMA1	1	X								
	2	X								
	3	X								
	4	X								
EMA2	1		X							
	2		X							
	3		X							
	4					X				
	5				X					
	6				X					
	7(a)							X		
	8		X							
	9		X							
EMA3	1			X						
	2			X						
	3			X						
	4			X						
	5						X			
	6						X			
	7					X				
	8						X			
	9(a)					X				
	10							X		
	11		X							
	12(a)		X			X				
	13		X							
	14		X			X				
	15		X			X				
	16		X							
	17		X							
	18		X							
EMA4	1					X				
	2					X				
	3					X				
	4					X				
	5					X				
	6					X				
	7					X				
	8				X					
	9				X					
	10				X					
	11				X					
	12				X					
	14				X					
	EMA5	1								X
2									X	
3									X	
4									X	
5							X		X	
EMA6	1									X
	2									X
	3									X
	4						X			X
	5						X			X
	6						X			X



Appendix B: Definition of variables used in the analysis

The variables that are used in the analysis are defined in Tables B1 – B4.

Table B1: EMS classification variables

VAR	Type	Description
M ₁	Indicator	EMS class 1 EMAS
M ₂	Indicator	EMS class 2 ISO 14001
M ₃	Indicator	EMS class 3 None

Table B2: Environmental management variables

VAR	Type	Description
X	Continuous	EMA Average
X ₁₁	Continuous	EMA11 Environmental policy
X ₁₂	Continuous	EMA12 Operational and risk management
X ₁₃	Continuous	EMA13 Commitment to training and awareness
X ₁₄	Continuous	EMA14 Compliance and conformance control
X ₁₅	Continuous	EMA15 Performance monitoring
X ₁₆	Continuous	EMA16 Open communication culture
X ₁₇	Continuous	EMA17 Documentation control
X ₁₈	Continuous	EMA18 Management review
X ₁₉	Continuous	EMA19 Reporting environmental performance

Table B3: Environmental outcome variables

VAR	Type	Description
Y ₁	Continuous	EP1 Compliance
Y ₂	Continuous	EP2 Conduct
Y ₃	Continuous	EP3 Process efficiency – raw materials
Y ₄	Continuous	EP4 Process efficiency - water
Y ₅	Continuous	EP5 Process efficiency - energy
Y ₆	Continuous	EP6 Releases - air
Y ₇	Continuous	EP7 Releases - water
Y ₈	Continuous	EP8 Releases - land



Table B4: Site characteristic variables

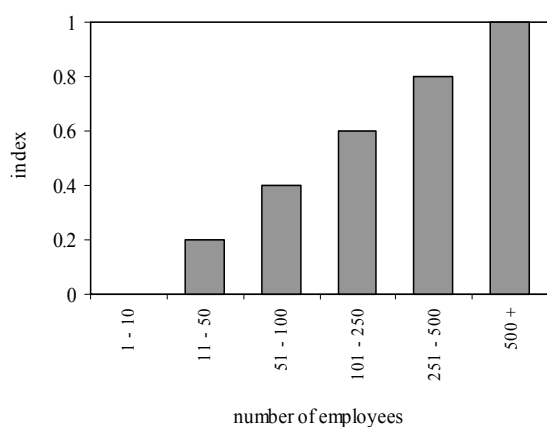
VAR	Type	Description
P	Index	Plant size
C	Index	Company size
S ₁	Indicator	Sector 1 Cement
S ₂	Indicator	Sector 2 Combustion
S ₃	Indicator	Sector 3 Inorganic chemicals
S ₄	Indicator	Sector 4 Organic chemicals
S ₅	Indicator	Sector 5 Paper and pulp
S ₆	Indicator	Sector 6 Other
S ₆₁	Indicator	Sector 5 or 6 Other (including paper and pulp)
S ₆₂	Indicator	Sector 2 or 6 Other (including combustion)

Indicator variables take the value one or zero, depending on whether the description applies to the site, or not. For example, if a site has ISO 14001 and operates in the organic chemicals sector, then $M_2 = S_4 = 1$ while $M_1 = M_3 = S_1 = S_2 = S_3 = S_5 = S_6 = 0$.

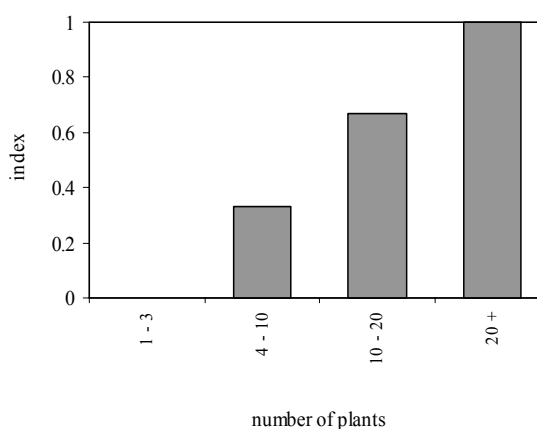
The indices for plant size and company size are based respectively on bands of employee numbers working at the plant, and plant numbers operated by the company. In both cases, the index is linear, and it is normalised so that it is equal to zero for the smallest band, and equal to one for the largest band (see Figure B1).

Figure B1: Construction of size indices

a) Plant size



b) Company size





Appendix C: Breakdown of sample sites with EP scores for process efficiency and releases

Table C1: Process efficiency – raw materials (EP3)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	8	3	2	13
Combustion	0	0	0	0
Inorganic chemicals	0	0	0	0
Organic chemicals	0	0	0	0
Paper & pulp	0	3	1	4
Other	0	2	2	4
Total	8	8	5	21

Table C2: Process efficiency – water (EP4)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	0	0	0	0
Combustion	0	0	0	0
Inorganic chemicals	0	0	0	0
Organic chemicals	0	0	0	0
Paper & pulp	0	3	2	5
Other	0	3	0	3
Total	0	6	2	8

Table C2: Process efficiency – energy (EP5)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	8	3	2	13
Combustion	0	1	0	1
Inorganic chemicals	0	0	0	0
Organic chemicals	0	0	0	0
Paper & pulp	0	3	2	5
Other	0	4	1	5
Total	8	11	5	24



d) Releases to air (EP6)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	8	3	2	13
Combustion	0	8	1	9
Inorganic chemicals	0	3	2	5
Organic chemicals	0	3	4	7
Paper & pulp	0	3	1	4
Other	0	6	2	8
Total	8	26	12	46

e) Releases to water (EP7)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	8	1	2	11
Combustion	0	5	1	6
Inorganic chemicals	0	3	2	5
Organic chemicals	0	4	5	9
Paper & pulp	0	3	2	5
Other	0	4	1	5
Total	8	20	13	41

f) Releases to land (EP8)

Industrial sector	EMS classification			Total
	EMAS	ISO 14001	No EMS	
Cement	0	0	0	0
Combustion	0	0	0	0
Inorganic chemicals	0	0	0	0
Organic chemicals	0	0	0	0
Paper & pulp	0	3	1	4
Other	0	4	1	5
Total	0	7	2	9



Appendix D: Stage 1 analysis

The first stage of the analysis considers the relationship between EMS classification and operator performance. The underlying model for the analysis is:

$$X_i = \alpha_0 + \alpha_1 M_{1i} + \alpha_3 M_{3i} + \upsilon_i \quad \dots (1)$$

$$S_{ji} = \beta' V_{ji} + \varepsilon_{ji} \quad \dots (2)$$

$$M_{ji} = \begin{cases} 1 & \text{if } U_{ji} > \text{Max } [S_{1i}, S_{2i}, S_{3i}] \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2, 3 \quad \dots (3)$$

The model comprises two structural equations and a decision rule. In equation (1), operator performance (X_i) is determined by the site's EMS type and a random disturbance term (υ_i). The value of the constant term (α_0) gives the expected operator performance score of an ISO 14001 site; while the coefficients of the two indicator variables (M_{1i} and M_{3i}) give the deviations in expected performance from this reference level for an EMAS site and a no-EMS site respectively. The *a priori* expectation is that the first of these coefficients is positive, and that the second is negative.

In equation (2), the dependent variable U_{ji} provides an index measure of the (unobserved) "sentiment" of the management of site i towards EMS type j ; while the vector V_{ji} denotes the (observed) attributes of EMS type j for that site. Thus the model allows for the possibility the attributes of each EMS type may vary between sites. This might be the case, for example, if one of the attributes is the associated marketing benefit, which may depend on the sector in which the site operates. Equation (3) defines the decision rule. Thus, it is assumed that management chooses the EMS type for which it has the highest sentiment; with the choice being reflected in the values of the indicator variables for the different types.

There is no reason to expect that there is any correlation of the disturbance terms between sites. However, given the similarities of the two types of EMS, it seems likely that if the management of a particular site has an above average sentiment towards one type, it will also have an above average sentiment towards the other. Furthermore, if this reflects a general attitude towards environmental management, then it is likely that the site will have a higher than average level of operator performance. If this is the case, then the disturbance terms υ_i , ε_{1i} and ε_{2i} are positively correlated within sites. To allow for this possibility, it is assumed that the disturbance terms in equation (2) follow a generalised extreme-value distribution of the form:

$$F(\boldsymbol{\varepsilon}) = \exp \left[- \sum_{i=1}^n G^i(\Psi_{1i}, \Psi_{2i}, \Psi_{3i}) \right]$$

$$\text{where } G^i(\cdot) = \left[\Psi_{1i}^\rho + \Psi_{2i}^\rho \right]^{1-\sigma} + \Psi_{3i} \quad \rho = 1 / (1 - \sigma)$$

$$\Psi_{ji} = \exp(-\varepsilon_{ji})$$



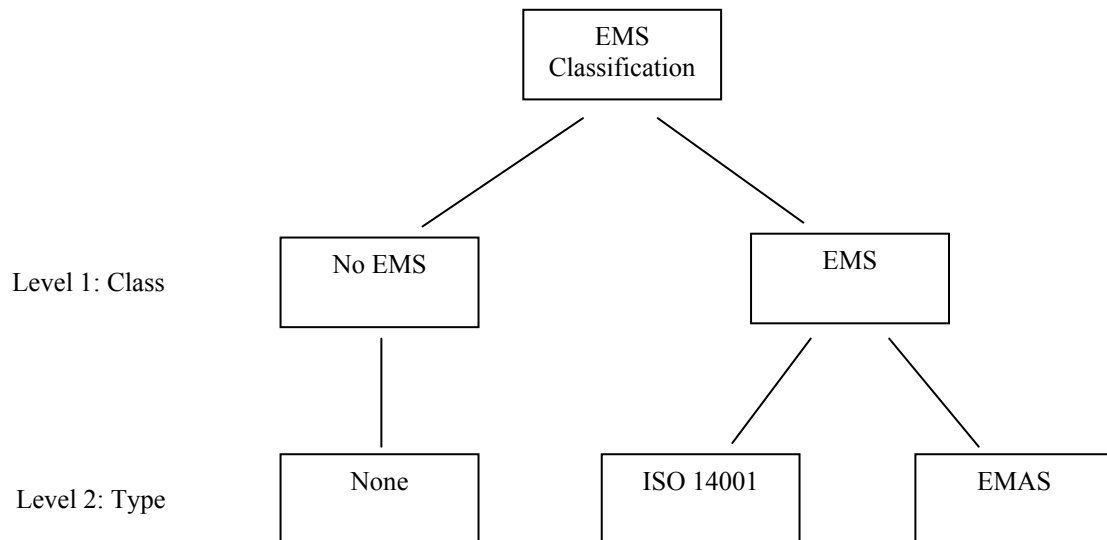
The value of the parameter σ lies between zero and one; being equal to zero if there is no correlation between the disturbance terms.

Because of the potential correlation between the disturbance terms, equation (1) must be estimated in a two-step procedure. In the first step, the probabilities of each site / type combination (denoted by $M_{ji}^{\#}$) are estimated using a probabilistic choice model. In the second step, these probabilities are used as instrumental variables (in place of the indicator variables M_{ji}) in the estimation of equation (1) by ordinary least squares.

Step 1: Choice of EMS type

Given the assumed distribution of the disturbance terms in equation (2), the values of the coefficients (β) can be estimated using a nested logit model. In this model the three EMS types are divided into two classes – as shown in Figure D1. It should be noted that the structure of the model reflects the assumptions that have been made regarding the correlation of the disturbance terms. It does not (necessarily) represent the actual decision process that is used for choosing a site’s EMS type.

Figure D1: Structure of nested logit model



The attributes of the different EMS types are captured by the use of indicator variables (EMAS, ISO and NONE). These are interacted with the indicator variables for sector (at the type level), and with the index variables for plant size and company size (at the class level). This allows for the possibility that management sentiment towards EMS in general may be affected by the size of the plant or the parent company, and that the relative attractiveness of the two types of EMS may differ between sectors.

The sample coefficient values for the nested logit model are given in Table D1.²⁴ The values of the coefficients for the level 2 (type) variables are all negative. There is no direct interpretation of these values. However, for each sector the relative values of the coefficients for the two types of EMS system determine which of the two has the greatest probability of being chosen – conditional on one of them being chosen (i.e. that class = EMS). If the value of the EMAS coefficient is higher (i.e. the magnitude of the negative value is smaller), then EMAS is relatively more attractive than ISO 14001, and hence is more likely to be chosen. This is the case for sector 1 only (where $-2.261 > -3.242$). For all other sectors, it is more likely that ISO 14001 will be chosen. The values of the coefficients for the two level 1 (class) variables are both positive, although only the plant size parameter is significant. This implies that management sentiment towards EMS (of both types) is directly related to plant size, and hence that larger plants are more likely to adopt an EMS.

Table D1: Parameter estimates for nested logit model

Dependent variable: EMS choice

	Coeff.	(P-value)*
<i>Type</i>		
ISO * S ₁	-3.242	(0.00)
ISO * S ₂	-0.341	(0.84)
ISO * S ₃	-3.786	(0.05)
ISO * S ₄	-4.504	(0.00)
ISO * S ₅	-5.674	(0.00)
ISO * S ₆	-3.934	(0.00)
EMAS * S ₁	-2.261	(0.00)
EMAS * S ₂	-35.607	(.)
EMAS * S ₃	-36.492	(.)
EMAS * S ₄	-6.114	(0.00)
EMAS * S ₅	-36.638	(.)
EMAS * S ₆	-36.474	(.)
<i>Class</i>		
EMS * P	3.974	(0.00)
EMS * C	0.600	(0.68)
<i>Inclusive value parameters</i>		
EMS	0.544	(0.00)
NO-EMS	1.000	(.)
Number of observations	171	
Number of groups	57	
LR	51.3	(0.00)

* P-values are based on outer-product gradient (OPG) standard errors.

²⁴ Coefficients and test statistics that are significant at the 10% level are highlighted with grey shading. This convention is used in all of the results tables in Appendix D and Appendix E.



Finally, the values of the coefficients for the inclusive value parameters for each class are shown. The coefficient of the NO-EMS class is constrained to be equal to one – reflecting the assumed distribution of the disturbance terms. The coefficient of the EMS parameter provides an estimate of the distribution parameter $(1 - \sigma)$. Since this is significantly different from one²⁵, it implies that the disturbance terms ε_{1i} and ε_{2i} are indeed correlated.

The estimated model is used to calculate the probabilities of the different EMS types being chosen for each site, and hence to predict which type will be chosen.²⁶ Table D2 compares the actual number of sites of each EMS type with the number predicted by the model. The figures in the final column are the actual number of sites of each type, while the figures in the final row are the predicted number. The figures in the cells are the number of sites of actual type A with predicted type B. Thus, for example, three sites with ISO 14001 are predicted to have EMAS, and so on. The figure highlighted in grey represent the number of correct predictions. Overall, the model predicts the correct EMS type for 72% of the sites, and the normalised prediction success index is 0.55.²⁷

Table D2: Prediction success table for EMS type

		Predicted type			Total
		EMAS	ISO 14001	None	
Actual type	EMAS	8	1	0	9
	ISO 14001	3	23	5	31
	None	2	5	10	17
Total		13	29	15	57

Step 2: Impact on operator performance

Having obtained the predicted probabilities of the three different EMS types for each site, these can be used as instrumental variables in the estimation of equation (1) by ordinary least squares. Ten different versions of the equation are estimated. The first version relates to the impact of EMS type on average operator performance (X); the other nine to the impact on the individual dimensions of performance ($X_{11} - X_{19}$).

In each case, the model specification is tested for omitted variables using the Ramsey RESET test, and for heteroskedasticity using a Breusch-Pagan test. The omitted variable test is only significant for two of the versions – those relating to *operational and risk*

²⁵ This cannot be inferred from the information provided in Table B1. However, the 95% confidence interval for the coefficient of the EMS IV parameter is [0.354, 0.734].

²⁶ The predicted choice is the one with the highest probability. For example, if the choice probabilities for EMAS, ISO 14001 and NONE are 0.2, 0.5 and 0.3 respectively, then the predicted choice is ISO 14001.

²⁷ The prediction success index was proposed by Mc Fadden *et al* (1977) as a goodness-of-fit measure for probabilistic-choice models. It is equal to $\sum_m \rho_m (\sigma_m - \rho_m)$, where ρ_m is the proportion of sites predicted to be of type m , and σ_m is the proportion of predicted type m sites that are correct.



management (X_{12}), and to *reporting environmental performance* (X_{19}). In contrast, the test for heteroskedasticity is significant for seven of the ten versions.²⁸ For these versions, White-corrected standard errors are used to calculate the reported p-values. In these cases, the adjusted R^2 value and the estimator s are not available, and the F test for the joint significance of the coefficients is replaced by a Wald test, where the test statistic is calculated using the robust covariance matrix.

The estimated parameter values for the version of the equation relating to average operator performance (X) are shown in Figure D3. As can be seen, the coefficients of the two instrumental variables have the expected signs, and both are strongly significant. A Hausman test of the difference between the two coefficient values and those obtained using the corresponding indicator variables (M_1 and M_3) is significant at the 10% level; justifying the use of the probabilities as instrumental variables. Interestingly, the magnitudes of the two coefficients are greater when the probabilities are used, which implies that – all else being equal – sites with below average operator performance are more likely to choose to implement an EMS.

Table D3: Parameter estimates using predicted probabilities as instrumental variables

Dependent variable: Overall operator performance (EMA)

X	Coeff.	(P-value)
Constant	1.701	(0.00)
$M_1^{\#}$	0.220	(0.00)
$M_3^{\#}$	-0.258	(0.00)
n	57	
R^2	0.41	
F	28.6	(0.00)
s	n/a	

The estimated parameter values for the equations relating to the individual dimensions of operator performance are shown in Table D4. As can be seen, the picture is broadly the same as for average operator performance. The two exceptions are the equation relating to *operational and risk management* (X_{12}) – in which the EMAS coefficient is negative (but insignificant), and the equation relating to *commitment to training and awareness* (X_{13}) – in which the coefficients are individually and jointly insignificant. In the first case, the unexpected value of the coefficient may be due to omitted variable bias, as the Ramsey RESET test is significant for this equation. However, in the second case, the implication is that neither type of EMS has an impact on this dimension of performance.

²⁸The Breusch-Pagan test for heteroskedasticity was conducted using the fitted values of the dependent variable. The test statistic was significant (at the 10% level) for X , X_{11} , X_{12} , X_{14} , X_{15} , X_{16} and X_{18} . The Ramsey RESET test for omitted variables was significant (at the 10% level) for X_{12} and X_{19} .



Table D4: Parameter estimates using predicted probabilities as instrumental variables (p-values in brackets)

Dependent variables: Individual dimensions of operator performance

	X ₁₁	X ₁₂	X ₁₃	X ₁₄	X ₁₅	X ₁₆	X ₁₇	X ₁₈	X ₁₉
Constant	1.770 (0.00)	1.769 (0.00)	1.611 (0.00)	1.783 (0.00)	1.669 (0.00)	1.539 (0.00)	1.751 (0.00)	1.687 (0.00)	1.584 (0.00)
M ₁ [#]	0.170 (0.01)	-0.002 (0.52)	0.147 (0.12)	0.149 (0.01)	0.296 (0.00)	0.451 (0.00)	0.105 (0.13)	0.325 (0.00)	0.464 (0.00)
M ₃ [#]	-0.206 (0.02)	-0.333 (0.00)	-0.007 (0.48)	-0.411 (0.00)	-0.173 (0.03)	-0.303 (0.01)	-0.388 (0.00)	-0.345 (0.01)	-0.292 (0.01)
n	57	57	57	57	57	57	57	57	57
R ²	0.22	0.21	0.03	0.29	0.40	0.41	0.27	0.27	0.40
F	9.4 (0.00)	6.6 (0.07)	0.9 (0.42)	14.7 (0.00)	37.9 (0.00)	42.9 (0.00)	10.1 (0.00)	12.6 (0.00)	17.8 (0.00)
s	n/a	n/a	0.22	n/a	n/a	n/a	0.22	n/a	0.19

In all of the other seven equations the coefficients have the expected sign, and all but one are significant. For these equations, the impact of ISO 14001 on performance ranges from 0.17 to 0.41; being greatest for *compliance and conformance control* (X₁₄) and for *documentation control* (X₁₇), and least for *environmental policy* (X₁₁) and for *performance monitoring* (X₁₅). The incremental impact of EMAS ranges from 0.10 to 0.46; being greatest for *open communication culture* (X₁₆) and for *reporting environmental performance* (X₁₉), and least for the two dimensions along which ISO 14001 has the greatest impact. However, it is possible that the incremental impact of EMAS on reporting environmental performance is overstated, as the omitted variable test is significant for this equation.



Appendix E: Stage 2 analysis

The second stage of the analysis considers the relationship between operator performance and various measures of environmental performance. The underlying model for the analysis is:

$$Y_i = F'(X_i, Z_i) + v_i \quad \dots (4)$$

That is, the environmental performance for a site (Y_i) is a function of its environmental management scores $\mathbf{X}_i = (X_{11i}, \dots, X_{19i})$, its characteristics $\mathbf{Z}_i = (P_i, C_i, S_{1i}, \dots, S_{6i})$, and a random disturbance term v_i . It is assumed that the disturbance terms are independent across sites, and follow a normal distribution with zero mean and constant variance.

Because the algorithms that have been used for calculating the environmental performance scores are sector-specific (apart from the measures relating to regulation and conduct), it seems plausible that the relationship between operator performance and environmental performance may vary from sector to sector. One way of allowing for this possibility is to estimate separate equations for each sector. However, given the small number of observations for each sector (i.e. 5 – 13), the standard errors of the coefficient estimates will be large. An alternative approach is to estimate a “pooled” equation that includes interaction terms $S_1X_{11}, S_2X_{11}, \dots, S_6X_{19}$. If the coefficients of these “artificial” variables are significantly different from zero, then one can conclude that the relationship does vary between sectors.

Unfortunately, it is not possible to do this for all nine dimensions of environmental management, as this would increase the number of independent variables in the model to seventy-two (including the constant term), making it impossible to estimate from this sample data. Consequently, two versions of equation (4) are used for the analysis. In the first version, the site’s overall environmental management score replaces the vector of individual scores, but the impact is allowed to vary across sectors. This reflects an underlying assumption that the relative values of the sector-adjusted coefficients of the individual dimensions of environmental management are equal to their respective weights in the calculation of the average management score.²⁹ In the second version, all of the individual dimensions of environmental management are included, but no allowance is made for differences in the relationship between sectors. Underlying this version is an assumption that the relationship between environmental management and outcomes does not vary between sectors.

Impact of overall environmental management quality

The relationship between environmental performance and overall environmental management is assessed using the following linear version of equation (4)

$$Y_i = \alpha_0 + \alpha_1' \mathbf{Z}_i + \alpha_2' \tilde{\mathbf{X}}_i + v_i \quad \dots (4a)$$

where $\tilde{\mathbf{X}}_i = (X_i, S_{1i}X_i, S_{2i}X_i, \dots, S_{6i}X_i)$

²⁹ These are equal to the proportion of the total number of questions that apply to each dimension.



The equation is estimated by ordinary least squares for each of the six EP measures for which there is sufficient data. By default, all of the relevant sectors are included in a single pooled equation. However, for three of the measures (EP2, EP3 and EP5) there is evidence to suggest that the variance of the disturbance differs between sectors.³⁰ Consequently, for these measures, the sectors are divided into two pools, and separate equations are estimated for each pool.

The sample coefficient values for these pooled equations are shown in Tables E1-E6, together with summaries of the implied sector-specific parameter values for equation (4). Also shown are the R^2 value for the fitted equation; the F statistic for the joint significance of the explanatory variables; and the sample value of the estimator for the variance of the disturbance term (s). In most cases coefficient values are provided for two regression equations. In the “full” equation, indicator and interaction variables are included for all of the sectors in the pool.³¹ In the “reduced” equation, sector variables with insignificant coefficient values (at the 10% level) are omitted.³² The values shown in the summary tables are equal to – or derived from – the coefficient values for the reduced equations.

As with the estimation of equation (2), the model specification is tested for omitted variables using the Ramsey RESET test, and for heteroskedasticity using a Breusch-Pagan test. The omitted variable test is significant for only one equation – the reduced pool B equation for *conduct* (Y_2). The test for heteroskedasticity is significant for three equations – the two reduced equations for *conduct* (Y_2), and the reduced pool A equation for *process efficiency – raw materials* (Y_3). For these three equations, White-corrected standard errors are used to calculate the reported p-values, and the F test for the joint significance of the coefficients is replaced by a Wald test.

It is clear from Table E1 and Table E2 that the sample provides no evidence to support the hypothesis that higher levels of operator performance lead to better *compliance* (EP1), or to better *conduct* (EP2). None of the sample values of the EMA coefficients for these measures are significant at the 10% level, and all but one are negative.

For the other four EP measures, the picture is mixed. For the two process efficiency measures (Table E3 and Table E4), four of the five coefficient values are positive, although only one is significant. However, it should be noted that because of the need to split the sectors into two pools, the sample sizes on which these calculations are based are very small – ranging from 5 to 19. For the two releases measures (Table E5 and E6), five out of eleven coefficient values are positive. However, all of these are significant.

³⁰ This is based on a comparison of the mean square error values of the calculated regression equations for each individual sector (not shown).

³¹ One sector is always omitted from the equation in order to avoid perfect co-linearity between the explanatory variables.

³² Insignificant sector variables are eliminated iteratively (starting with the least significant – i.e. those with the highest P-values), until all those that remain are significant at the 10% level on a two-tail test.



Table E1: Regulation (EP1)

a) Parameter estimates for pooled equation

Y ₁	Coeff.	(PV)	Coeff.	(PV)
Const	1.810	(0.01)	2.081	(0.00)
P	-0.002	(0.97)	-0.011	(0.87)
C	-0.005	(0.93)	-0.015	(0.78)
X	0.036	(0.93)	-0.127	(0.86)
S ₁	0.542	(0.58)		
S ₁ X	-0.333	(0.55)		
S ₃	1.383	(0.12)	1.106	(0.07)
S ₃ X	-0.875	(0.11)	-0.702	(0.06)
S ₄	0.685	(0.38)		
S ₄ X	-0.435	(0.36)		
S ₅	-0.911	(0.33)		
S ₅ X	0.570	(0.32)		
S ₆	-0.069	(0.93)		
S ₆ X	0.044	(0.93)		
n	57		57	
R ²	0.25		0.14	
F	1.1	(0.38)	1.6	(0.17)
s	0.12		0.11	

PV = P-Value

b) Summary of coefficient estimates by sector

Y ₁	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6
Const	2.081	2.081	3.187	2.081	2.081	2.081
P	-0.011	-0.011	-0.011	-0.011	-0.011	-0.011
C	-0.015	-0.015	-0.015	-0.015	-0.015	-0.015
X	-0.127	-0.127	-0.829	-0.127	-0.127	-0.127
n	57	57	57	57	57	57
R ²	0.14	0.14	0.14	0.14	0.14	0.14
F	1.6	1.6	1.6	1.6	1.6	1.6
s	0.11	0.11	0.11	0.11	0.11	0.11



Table E2: Conduct (EP2)

a) Parameter estimates for pooled equation

Pool A (sectors S1, S5)				Pool B (sectors S2, S3, S4, S6)			
Y ₂	Coeff.	(PV)	Coeff.	(PV)	Y ₂	Coeff.	(PV)
Const	3.241	(0.13)	3.829	(0.00)	Const	2.691	(0.00)
P	-0.848	(0.12)	-0.313	(0.26)	P	-0.035	(0.65)
C	0.336	(0.23)	0.406	(0.13)	C	-0.066	(0.33)
X	-0.742	(0.71)	-1.258	(0.97)	X	-0.417	(0.90)
S ₅	-1.753	(0.48)			S ₂	0.118	(0.88)
S ₅ X	1.289	(0.36)			S ₂ X	-0.055	(0.91)
					S ₄	-0.055	(0.93)
					S ₄ X	-0.000	(1.00)
					S ₆	0.215	(0.75)
					S ₆ X	-0.188	(0.65)
n	18		18		n	39	
R ²	0.52		0.28		R ²	0.40	
F	2.6	(0.08)	1.8	(0.20)	F	2.1	(0.06)
s	0.25		0.28		s	0.10	

PV = P-Value

b) Summary of parameter estimates by sector

Y ₂	Coeff.	(PV)	Sector 3	Sector 4	Sector 5	Sector 6
Const	3.241	2.594	2.594	2.594	1.488	2.594
P	-0.848	-0.058	-0.058	-0.058	-0.848	-0.058
C	0.336	-0.027	-0.027	-0.027	0.336	-0.027
X	-0.742	-0.385	-0.385	-0.385	0.547	-0.385
n	18	39	39	39	18	39
R ²	0.52	0.25	0.25	0.25	0.52	0.25
F	2.6	3.8	3.8	3.8	2.6	3.8
s	0.25	0.11	0.11	0.11	0.25	0.11



Table E3: Process efficiency – raw materials (EP3)

a) Parameter estimates for pooled equations

Pool A (sector S1)			Pool B (sector S7)		
Y ₃	Coeff.	(PV)	Y ₃	Coeff.	(PV)
Const	0.317	(0.82)	Const	0.841	(0.74)
P	-0.237	(0.52)	P	-0.773	(0.38)
C	0.154	(0.56)	C	0.398	(0.54)
X	0.617	(0.54)	X	0.731	(0.33)
n	13		n	8	
R ²	0.73		R ²	0.25	
F	8.3	(0.01)	F	0.4	(0.74)
s	0.08		s	0.43	

PV = P-Value

b) Summary of parameter estimates by sector

Y ₃	Sector 1	Sector 7
Const	0.317	0.841
P	-0.237	-0.773
C	0.154	0.398
X	0.617	0.731
n	13	8
R ²	0.73	0.25
F	8.3	0.4
s	0.08	0.43



Table E4: Process efficiency – energy (EP5)

a) Parameter estimates for pooled equations

Pool A (sectors S1, S8)			Pool B (sector S5)		
Y ₅	Coeff.	(PV)	Y ₅	Coeff.	(PV)
Const	-3.334	(0.12)	Const	-1.501	(0.79)
P	0.015	(0.97)	P	0.340	(0.95)
C	-0.504	(0.09)	C	-0.710	(0.69)
X	2.840	(0.02)	X	1.939	(0.32)
S ₈	5.301	(0.06)			
S ₈ X	-2.888	(0.07)			
n	19		n	5	
R ²	0.43		R ²	0.44	
F	2.0	(0.14)	F	0.3	(0.86)
s	0.26		s	0.75	

PV = P-Value

b) Summary of parameter estimates by sector

Y ₅	Sector 1	Sector 5	Sector 8
Const	-3.334	-1.501	1.967
P	0.015	0.340	0.015
C	-0.504	-0.710	-0.504
X	2.840	1.939	-0.048
n	19	5	19
R ²	0.43	0.44	0.43
F	2.0	0.3	2.0
s	0.26	0.75	0.26

Table E5: Releases to air (EP6)

a) Parameter estimates for pooled equation

Y_6	Coeff.	(PV)	Coeff.	(PV)
Const	0.769	(0.61)	1.971	(0.01)
P	0.190	(0.42)	-0.008	(0.97)
C	-0.277	(0.13)	-0.245	(0.17)
X	0.604	(0.26)	-0.127	(0.61)
S_1	-3.315	(0.16)	-4.558	(0.02)
$S_1 X$	1.655	(0.23)	2.464	(0.03)
S_2	-0.751	(0.75)		
$S_2 X$	0.404	(0.78)		
S_4	3.487	(0.11)		
$S_4 X$	-2.280	(0.10)		
S_7	2.048	(0.35)		
$S_7 X$	-1.376	(0.30)		
n	46		46	
R^2	0.32		0.18	
F	1.4	(0.20)	1.8	(0.13)
s	0.31		0.31	

PV = P-Value

b) Summary of parameter estimates by sector

Y_6	Sector 1	Sector 2	Sector 3	Sector 4	Sector 7
Const	-2.587	1.971	1.971	1.971	1.971
P	-0.008	-0.008	-0.008	-0.008	-0.008
C	-0.245	-0.245	-0.245	-0.245	-0.245
X	2.337	-0.127	-0.127	-0.127	-0.127
n	46	46	46	46	46
R^2	0.18	0.18	0.18	0.18	0.18
F	1.8	1.8	1.8	1.8	1.8
s	0.31	0.31	0.31	0.31	0.31

Table E6: Releases to water (EP7)

a) Parameter estimates for pooled equation

Y_7	Coeff.	(PV)	Coeff.	(PV)
Const	-1.143	(0.48)	-1.432	(0.17)
P	0.598	(0.04)	0.414	(0.09)
C	0.092	(0.69)	-0.028	(0.90)
X	1.274	(0.11)	1.542	(0.01)
S_1	5.732	(0.03)	5.265	(0.03)
$S_1 X$	-3.074	(0.05)	-2.809	(0.04)
S_2	-2.505	(0.42)		
$S_2 X$	1.533	(0.41)		
S_4	4.234	(0.05)	4.445	(0.01)
$S_4 X$	-2.747	(0.04)	-2.868	(0.01)
S_5	0.889	(0.72)		
$S_5 X$	-0.508	(0.74)		
S_6	-2.155	(0.52)		
$S_6 X$	1.088	(0.59)		
n	41		41	
R^2	0.58		0.48	
F	2.9	(0.01)	4.4	(0.00)
s	0.33		0.33	

PV = P-Value

b) Summary of parameter estimates by sector

Y_7	Sector 1	Sector 2	Sector 3	Sector 4	Sector 5	Sector 6
Const	3.833	-1.432	-1.432	3.013	-1.432	-1.432
P	0.414	0.414	0.414	0.414	0.414	0.414
C	-0.028	-0.028	-0.028	-0.028	-0.028	-0.028
X	-1.267	1.542	1.542	-1.326	1.542	1.542
n	41	41	41	41	41	41
R^2	0.48	0.48	0.48	0.48	0.48	0.48
F	4.4	4.4	4.4	4.4	4.4	4.4
s	0.33	0.33	0.33	0.33	0.33	0.33

Impact of individual dimensions of environmental management

The relationship between environmental performance and the individual dimensions of environmental management is assessed using the following linear version of equation (4)

$$Y_i = \beta_0 + \beta_1' \tilde{Z}_i + \beta_2' X_i + \omega_i \quad \dots (4b)$$

where $\tilde{Z}_i = (P_i, C_i)$

Again, the equation is estimated by ordinary least squares for each of the six EP measures for which there is sufficient data; with the model specification being tested for omitted variables using the Ramsey RESET test, and for heteroskedasticity using a Breusch-Pagan test in each case. The omitted variable test is significant for two equations – *conduct* (Y_2) and *process efficiency – raw materials* (Y_3). The test for heteroskedasticity is significant for the equation for *conduct* (Y_2). For this equation, White-corrected standard errors are used to calculate the reported p-values, and the F test for the joint significance of the coefficients is replaced by a Wald test.

The sample coefficient values for these equations are shown in Tables E7 – E12, together with the R^2 value for the fitted equation; the F statistic for the joint significance of the explanatory variables; and the sample value of the estimator for the variance of the disturbance term (s). For reference, coefficient values are also shown for an equation in which the overall management measure is used, but no allowance is made for any heterogeneity between sectors.

Table E7: Regulation (EP1)

Y_1	Coeff.	(P-value)	Coeff.	(P-value)
Const	2.190	(0.00)	2.159	(0.00)
P	-0.030	(0.65)	-0.031	(0.70)
C	n/a		n/a	
X	-0.192	(0.96)		
X_{11}			0.077	(0.35)
X_{12}			-0.043	(0.62)
X_{13}			-0.062	(0.76)
X_{14}			0.137	(0.22)
X_{15}			-0.004	(0.51)
X_{16}			-0.100	(0.80)
X_{17}			-0.071	(0.75)
X_{18}			-0.072	(0.69)
X_{19}			-0.054	(0.69)
n	57		57	
R^2	0.07		0.13	
F	2.0	(0.15)	0.7	(0.72)
s	0.11		0.12	



Table E8: Conduct (EP2)

Y ₂	Coeff.	(P-value)	Coeff.	(P-value)
Const	3.108	(0.00)	3.103	(0.00)
P	-0.151	(0.16)	-0.201	(0.13)
C	0.088	(0.44)	0.114	(0.38)
X	-0.731	(0.99)		
X ₁₁			-0.131	(0.65)
X ₁₂			0.439	(0.02)
X ₁₃			-0.074	(0.70)
X ₁₄			-0.324	(0.88)
X ₁₅			-0.399	(0.92)
X ₁₆			-0.113	(0.72)
X ₁₇			-0.135	(0.79)
X ₁₈			0.245	(0.13)
X ₁₉			-0.237	(0.87)
n	57		57	
R ²	0.25		0.38	
F	4.4	(0.01)	2.1	(0.04)
s	0.20		0.19	

Table E9: Process efficiency – raw materials (EP3)

Y ₃	Coeff.	(P-value)	Coeff.	(P-value)
Const	2.594	(0.03)	2.435	(0.24)
P	-0.038	(0.91)	-0.365	(0.49)
C	0.315	(0.20)	0.084	(0.75)
X	-0.732	(0.85)		
X ₁₁			1.820	(0.16)
X ₁₂			0.309	(0.36)
X ₁₃			-0.339	(0.72)
X ₁₄			-0.317	(0.65)
X ₁₅			-0.338	(0.67)
X ₁₆			0.796	(0.21)
X ₁₇			-0.999	(0.86)
X ₁₈			-0.850	(0.82)
X ₁₉			-0.654	(0.93)
n	21		21	
R ²	0.10		0.74	
F	0.6	(0.60)	2.3	(0.11)
s	0.29		0.22	

Table E10: Process efficiency – energy (EP5)

Y ₅	Coeff.	(P-value)	Coeff.	(P-value)
Const	0.145	(0.89)	-0.670	(0.72)
P	0.370	(0.32)	0.522	(0.40)
C	-0.326	(0.18)	-0.201	(0.47)
X	0.752	(0.12)		
X ₁₁			2.700	(0.02)
X ₁₂			-1.925	(0.97)
X ₁₃			0.416	(0.26)
X ₁₄			-0.216	(0.58)
X ₁₅			0.659	(0.19)
X ₁₆			0.563	(0.28)
X ₁₇			-1.290	(0.91)
X ₁₈			0.353	(0.36)
X ₁₉			-0.197	(0.64)
n	24		24	
R ²	0.14		0.60	
F	1.1	(0.36)	1.6	(0.21)
s	0.33		0.29	

Table E11: Releases to air (EP6)

Y ₆	Coeff.	(P-value)	Coeff.	(P-value)
Const	1.813	(0.00)	2.301	(0.02)
P	0.047	(0.83)	0.071	(0.79)
C	-0.094	(0.58)	-0.081	(0.65)
X	-0.132	(0.64)		
X ₁₁			-0.867	(0.88)
X ₁₂			0.563	(0.11)
X ₁₃			0.749	(0.04)
X ₁₄			0.474	(0.14)
X ₁₅			0.083	(0.40)
X ₁₆			-0.279	(0.70)
X ₁₇			0.240	(0.33)
X ₁₈			-0.953	(0.98)
X ₁₉			-0.444	(0.88)
n	46		46	
R ²	0.01		0.21	
F	0.20	(0.89)	0.81	(0.63)
s	0.33		0.33	



Table E12: Releases to water (EP7)

Y ₇	Coeff.	(P-value)	Coeff.	(P-value)
Constant	-0.445	(0.55)	-0.339	(0.75)
P	0.212	(0.43)	0.265	(0.38)
C	-0.087	(0.72)	-0.161	(0.51)
X	1.081	(0.02)		
X ₁₁			-0.471	(0.73)
X ₁₂			-0.340	(0.73)
X ₁₃			0.365	(0.20)
X ₁₄			-1.261	(0.99)
X ₁₅			0.072	(0.43)
X ₁₆			0.974	(0.06)
X ₁₇			1.522	(0.01)
X ₁₈			0.068	(0.45)
X ₁₉			0.118	(0.39)
n	41		41	
R ²	0.16		0.44	
F	2.4	(0.08)	2.0	(0.06)
s	0.40		0.37	

Looking across Tables E7 – E12, only six of the fifty-four coefficient values for the environmental management dimensions are significant at a 10% level, and more than half are negative. With so few significant coefficients, it is impossible to draw any meaningful conclusions regarding the differential impacts of the management dimensions. However, it is interesting to note the differences between the different types of environmental outcome. For example, the coefficients for *environmental policy* (X₁₁) are significant for both process efficiency measures (Y₃ and Y₅), but negative for both releases measures (Y₆ and Y₇). Conversely, the coefficients for *performance monitoring* (X₁₅) are positive for both releases measures (and significant for one), but negative for both process efficiency measures. This suggests that the relative impacts of the different dimensions may vary between the different types of environmental outcome.

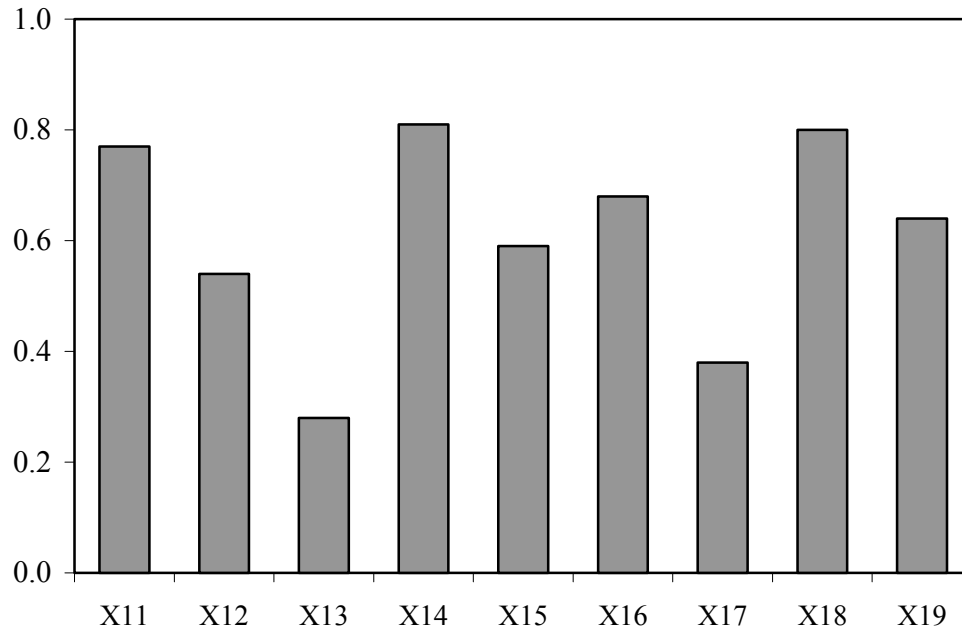
Despite the paucity of significant coefficient values, the R² values for the equations are not unreasonable for this type of data – ranging from 0.13 to 0.74, with an average of 0.42. This suggests that there may be *multicollinearity* between some, or all, of the explanatory variables. This is confirmed by the R² values for the “auxiliary regressions” for the individual management dimensions, which are shown in Figure E1.³³ It is clear from this that the degree of the problem varies between the different dimensions of environmental management. While the R² values are not particularly high for *commitment to training and awareness* (X₁₃), or for *documentation control* (X₁₇), it does appear that multicollinearity is a significant problem for *environmental policy* (X₁₁), *compliance and conformance control*

³³ In the auxiliary regressions, each management dimension is regressed against all of the other management dimensions and the two size-related variables. For example, in the auxiliary regression for X₁₁, the independent variables are P, C and X₁₂ – X₁₉.



(X_{14}) and *management review* (X_{18}). It is therefore somewhat surprising, that the first two of these dimensions account for half of the significant coefficients.

Figure E1: Auxiliary R^2 values for environmental management dimensions





Appendix F: Interpretation of regression results

This appendix provides a brief explanation of how the results of a statistical regression analysis for a given sample may be assessed, and of how general conclusions may be drawn from them.³⁴ There are various summary measures that can be used to do this; the most common of these being:

- R^2 value
- Adjusted R^2 value
- F-statistic
- t-statistic
- P-value

The first two measures relate to how well the model “fits” the sample data; while the last three are used to draw conclusions about the “true” values of the model parameters. Reflecting this division, the measures are considered under two separate headings – *goodness of fit*, and *hypothesis testing*. However, before this can be done, it is necessary to explain a few of the basic concepts of linear regression.

F1. Basic concepts

F1.1 The regression function

A key concept underlying the interpretation of regression results is the difference between the population regression function and the sample regression line.

The *population regression function* (PRF) gives the hypothesised statistical relationship between the dependent variable Y and a selected set of independent explanatory variables X_2, \dots, X_K , for the population as a whole. It is assumed that this relationship is linear, and is given by:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i \quad \text{for } i = 1, \dots, N$$

where $\beta_1 \dots \beta_K$ are the fixed (but unknown) coefficients of the independent variables; ε_i is a random disturbance term with mean (average) equal to zero and constant standard deviation (denoted by σ); and N is the population size.

For example, the dependent variable might be EP1; the independent explanatory variables might be EMA, plant size and company size; with the population being all sites regulated under IPPC.

The disturbance term can reflect a number of factors. It might reflect the fact that the dependent variable is affected by other variables, which have been omitted from the model; or that the dependent variable has been measured inaccurately; or that the value of the

³⁴ By necessity, this appendix provides only a cursory introduction to the topic of regression analysis. For those readers that wish to go into more detail, *A Guide to Econometrics* (5th Edition) by Peter Kennedy (Blackwell Publishing) is a comprehensive – but reasonably accessible – text.



dependent variable is inherently random. Clearly, when specifying the PRF one would like this disturbance term to be relatively small, and hence that the independent variables provide a good explanation for any variations in the values of the dependent variable between different members of the population.

If one had data for every member of the population (e.g. for all IPPC sites), then one could estimate the values of $\beta_1 \dots \beta_K$ directly. Of course, in practice this is rarely the case, and one only has data for a particular sample of observations. The *sample regression function* (SRF) gives the “fitted” relationship between the dependent variable Y and the selected explanatory variables for this sample. It takes the form.

$$Y_i = b_1 + b_2 X_{2i} + \dots + b_K X_{Ki} + e_i \quad \text{for } i = 1, \dots, n$$

where b_1, \dots, b_K are calculated from the sample data; e_i is the residual (or error) term for that observation; and $n (< N)$ is the sample size.

The first K terms of the SRF (i.e. omitting the error term) constitute the *sample regression line* (SRL). This gives the predicted value of the dependent variable for that observation (denoted by \hat{Y}_i). That is:

$$\hat{Y}_i = b_1 + b_2 X_{2i} + \dots + b_K X_{Ki} \quad \text{for } i = 1, \dots, n$$

By definition, the error term (or *residual*) for an observation is equal to the difference between the actual and predicted values of the dependent variable. That is:

$$e_i = Y_i - \hat{Y}_i \quad \text{for } i = 1, \dots, n$$

There are a number of different techniques that can be used to “fit” the SRL to the sample data. The most common technique – and the one that has been used for most of the analysis in this report – is called Ordinary Least Squares (OLS). Under this technique, the values of the coefficients are chosen so as to minimize the sum of the squared error terms (i.e. minimize $e_1^2 + \dots + e_n^2$).

F1.2 Estimators

The values b_1, \dots, b_K that are calculated from the sample data are known as the *estimators* for unknown population coefficients $\beta_1 \dots \beta_K$. Because they are calculated from the sample data, they will vary from sample to sample. That is, unlike the population coefficients which are the fixed, each SRF coefficient b_k is a random variable, with mean $E(b_k)$ and standard deviation $se(b_k)$; where the latter provides a measure of the “accuracy” of the estimator. It can be shown that provided certain conditions are met, then:

$$E(b_k) = \beta_k.$$

$$se(b_k) = \left(\frac{\sigma}{se(X_k)} \right) \left(\frac{1}{\sqrt{n-1}} \right) \left(\frac{1}{\sqrt{1-R_k^2}} \right)$$



where σ is the standard deviation of the disturbance term in the PRF; $se(X_k)$ is the standard error of the sample values of explanatory variable X_k ; and R_k^2 is the multiple coefficient of determination for the regression of variable X_k on all of the other explanatory variables (see below for an explanation of R^2).

In practice, the value of σ is unknown. Consequently, it is replaced by an unbiased estimator s , which is calculated from the error terms in the SRF:

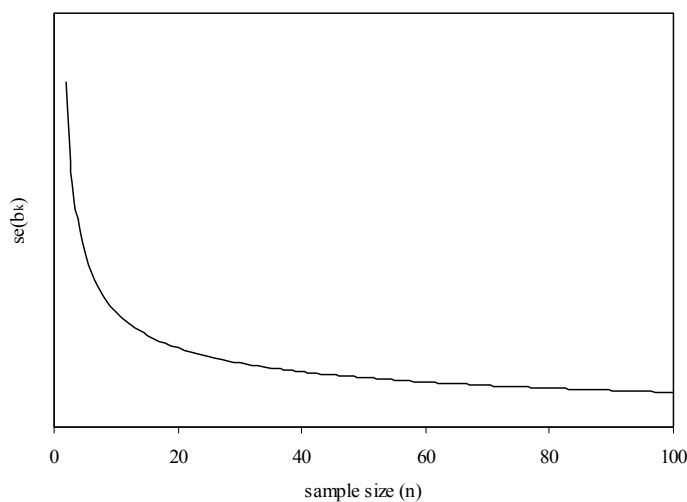
$$s = \sqrt{\frac{e_1^2 + e_2^2 + \dots + e_n^2}{n - K}}$$

There are two main points to be noted from the above expressions.

- If many different samples are taken, the average value of a calculated SRF coefficient over all the samples will be equal to the respective coefficient in the PRF.
- Assuming that σ (or s) is constant, the standard deviation of a coefficient estimator is affected by three factors: the degree of variation in the sample values for the variable; the size of the sample; and the degree of correlation between the variable and the other explanatory variables.

All else being equal, the standard deviation of an estimator will decline as the sample size increases. However, as can be seen in Figure F1, once one goes beyond around 40 observations, the rate of decrease flattens off considerably.

Figure F1: Impact of sample size on standard error of estimator





F1.3 Decomposition of the variation in the dependent variable Y

Having calculated the SRF, there are two types of question that one might want to answer.

- How well does the calculated SRL “fit” the data for the sample?
- What conclusions can be drawn about the PRF from the SRF?

As will be seen, there are a number of measures that can be used to answer these questions. However, all of these measures are based (in one way or another) on a decomposition of the variation in the sample values of the dependent variable into two components – one representing the variation that is explained by the SRL, the other representing the unexplained (or residual) variation.

If \tilde{Y} denotes the mean (average) value of the dependent variable across all observations in the sample, then it can be shown that:

$$\sum_{i=1}^n (Y_i - \tilde{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \tilde{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

SST = SSR + SSE

That is, the sum (across all sample observations) of the squared deviations of the actual values of the dependent variable from its mean value (SST) is equal to the sum of squared deviations of the predicted values from the mean value (SSR), plus the sum of squared deviations of the actual values from the predicted values, i.e. the sum of squared error terms (SSE). The first component represents the variation in the dependent variable that is explained by the model; while the second component represents the unexplained (or random) variation.

This decomposition is usefully summarized in a so-called *Analysis of Variance (ANOVA) Table*, which also includes the degrees of freedom³⁵ for each component, and the resultant “mean square variation”. It should be noted that while the variations are additive, the mean square variations are not (i.e. MST ≠ MSR + MSE).

Table F1: ANOVA Table for dependent variable

Source of variation	Variation	Degrees of Freedom	Mean square
Model	SSR	K – 1	MSR = SSR / (K – 1)
Error	SSE	n – K	MSE = SSE / (n – K)
Total	SST	n – 1	MST = SST / (n – 1)

³⁵ The degrees of freedom for a measure is equal to the number of independent observations on which the measure is based.



The Mean Square Error (MSE) is equal to the sum of squared error terms, divided by the degrees of freedom ($n - K$). Thus, the square root of the MSE (or *Root MSE*) is equal to s , which, as was noted above, is an unbiased estimator of the standard deviation of the random term in the PRF.

F2. Goodness of fit of the SRL

For OLS regressions, there are two measures that are commonly used to assess how well the SRL fits the sample data.

a) R^2 : Multiple coefficient of determination

This summary measure shows the proportion of the sample variation in the dependent variable Y that is explained by the SRL. As such, it provides a measure of the correlation between the actual values of the dependent variable and the values predicted by the SRL. It is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The higher the value of R^2 , the better the SRL fits the data. If $R^2 = 1$ then there is a perfect fit. However, it should be noted that the R^2 value will always increase (or at least not decrease) if additional variables are added to the model – even if these are not significant (i.e. the coefficients in the PRF are equal to zero).

b) $Adj. R^2$: Adjusted multiple coefficient of determination

Again this provides a measure of the extent to which the variation in the dependent variable that is explained by the model. However, unlike R^2 , it is based on the mean square components of the ANOVA. As such, it takes some account of the number of independent variables that have been included in the model. It is given by:

$$Adj. R^2 = 1 - \frac{MSE}{MST} = 1 - (1 - R^2) \left(\frac{n-1}{n-K} \right)$$

When the sample size is large relative to the number of variables (i.e. $n \gg K$), there will be very little difference between the R^2 value and the adjusted R^2 value. However, when $n - K$ is relatively small (as is the case in this analysis), there can be a considerable divergence between the two measures. Indeed, it is possible for the adjusted R^2 value to be negative. As with R^2 , higher values indicate a better fit to the sample data.

There is no absolute benchmark that can be used to say whether a particular R^2 value (or adjusted R^2 value) is high or low. To a certain extent, it depends on the type of data that is being analysed. A value of 0.3 may be high in some contexts, while a value of 0.8 may be low in others. With cross-sectional data (as is the case in this analysis), low R^2 values are relatively common.



F3. Hypothesis testing

While the R^2 value tells us something about the fit of the SRL to the sample data, it does not – in itself – allow us to conclude anything about the PRF. In order to do this, it is necessary to conduct formal statistical tests using the various results of the regression analysis.

The basic objective of hypothesis testing is to determine whether the sample observations are compatible with some prior belief, or not. The details of the test procedure will depend on the nature of the hypothesis that is being tested (i.e. the particular question that is being posed about the PRF). However, the general structure is the same in all cases. Given a *stated hypothesis* about the PRF, an appropriate *test statistic* is calculated from the SRF results, and this is compared with a *critical value*. If the value of the test statistic is greater than the critical value, then the hypothesis is rejected; if it is less than the critical value, then it is not. These three steps are considered in turn.

F3.1 Defining the hypothesis

The first step is to define the hypothesis that is to be tested. This is known as the *null hypothesis*, and is denoted by H_0 . This is tested against an *alternative hypothesis*, which is denoted by H_1 . In most cases, the alternative hypothesis is just the converse of the null hypothesis. There are many potential definitions for the null and alternative hypotheses, depending on the particular question that one is seeking to answer. Some of these can be relatively complex. However, for a basic assessment of the PRF there are two main questions that one would like to answer.

- Do the independent variables collectively have any impact on the dependent variable?
- What is the impact of a particular independent variable on the dependent variable?

These questions are considered in turn.

a) *The collective impact of all explanatory variables*

An obvious question to ask is whether the selected explanatory variables have any collective impact on the dependent variable. Clearly, a prerequisite for this is that at least some of the PRF coefficients β_2, \dots, β_K must be non-zero. Consequently, the appropriate null and alternative hypotheses are:

$$H_0: \beta_2 = \beta_3 = \dots = \beta_K = 0$$

H_1 : Not all of the coefficients are equal to zero

If the null hypothesis is not rejected in the test, then this implies that the model hypothesised in the PRF does no better than an alternative, “trivial” model that contains only a constant and a random error term.



b) *The individual impact of a particular explanatory variable*

Having established that the explanatory variables have a collective impact on the dependent variable, the next question is whether a particular explanatory variable has any impact. Indeed in most studies – including this one – this is the real matter of interest. For this question, the hypotheses can take two alternative forms.

Form (i)

$$H_0: \beta_k = \beta_k^*$$

$$H_1: \beta_k \neq \beta_k^*$$

Form (ii)

$$H_0: \beta_k \leq \beta_k^*$$

$$H_1: \beta_k > \beta_k^*$$

In the first case, one is asking whether the coefficient β_k differs from some pre-defined value (denoted by β_k^*). This is known as a two-sided (or a two-tail) test, reflecting the fact that it can be greater than, or less than this value. In the second case, one is asking whether the coefficient exceeds the pre-defined value. This is known as a one-sided (or one-tail) test.

The choice between a one-sided test and a two-sided test depends on whether one has any *a priori* expectations about the sign of the coefficient. In this analysis, a one-sided hypothesis seems more appropriate for testing the coefficients of the EMA scores, since one would expect these to be positive. In contrast, a two-sided hypothesis has been used for the size-related variables and the sector dummies, as there is no reason to expect the coefficients of these variables to have any particular sign.

For both forms of the hypothesis, the default is to set $\beta_k^* = 0$, and hence to test whether the coefficient β_k is different from (or greater than) zero. However, this does not have to be the case, and there may be situations where one wants to test whether the coefficient differs from some expected value.

There are two important points to note from these various hypotheses. The first point is that they all relate to the PRF. The second is that the tests are set up so that the prior belief is represented by the alternative hypothesis (H_1). This is because rejection of the null hypothesis implies that the alternative is true (with a given level of error), but acceptance of the null hypothesis implies only that it may be true (again with a given level of error). There may be other null hypotheses that would also not be rejected.

In deciding whether to reject the null hypothesis or not, there are two types of error that can be committed:

- the null hypothesis (H_0) may be rejected when it is, in fact, true;
- the null hypothesis (H_0) may not be rejected when it is, in fact, false.

The first is called a *type I error*, while the second is called a *type II error*. Ideally, one would want to minimize both types of error – i.e. only reject a null hypothesis when it is false, and only accept it when it is true. Unfortunately, for any given sample it is not possible to minimize both types of error simultaneously. In practice, it is assumed that committing a type I error is likely to be more serious, and therefore an upper limit is set for



the probability of its occurrence. The probability of the type II error is then minimized subject to this constraint.

This upper limit for the probability of a type I error occurring is called the *level of significance* for the test (and is denoted by α). Thus, if one sets $\alpha = 0.05$ then one is accepting that there is a 5% chance that the null hypothesis will be rejected when it is, in fact, true. For example, if the null hypothesis is $H_0: \beta_k \leq 0$, then the chance of incorrectly concluding that $\beta_k > 0$ is equal to 5%.

The choice of the appropriate value for α depends on a number of factors. The most important of these is the severity of the consequences of committing a type I error. Clearly, the more serious the negative consequences of incorrectly rejecting the null hypothesis, the smaller the value should be. A second factor to take into account is the size of the sample. For a given level of significance, the probability of committing a type II error declines as the sample size increases. Consequently, it may be desirable to take advantage of this, and reduce the value for α . For large sample sizes with critical consequences, a limit value of 0.01 (or less) is likely to be appropriate. For small sample sizes with less critical consequences a limit value of 0.10 may be more reasonable – and this is the value that has been assumed for this analysis.

F3.2 Computing the test statistic

Having defined the hypothesis that is to be tested, the next step is to compute an appropriate test statistic. The choice of test statistic depends on the nature of the hypothesis. Consequently, the two cases identified above are considered in turn.

a) *Collective impact of all explanatory variables*

In this case the appropriate test statistic is the F-statistic (denoted by F). It is defined as:

$$F = \frac{\sqrt{\frac{MSR}{MSE}}}{\sqrt{\frac{SSE}{n-K}}} = \frac{\sqrt{\left(\frac{SST - SSE}{K-1}\right) / \frac{SSE}{n-K}}}{\sqrt{\frac{SSE}{n-K}}} = \frac{R^2}{1-R^2} \left(\frac{n-K}{K-1} \right)$$

Because the values of MSR and MSE will vary from sample to sample, the F-statistic is a random variable. It is distributed according to the F probability distribution, with $(K-1, n-K)$ degrees of freedom.³⁶

As can be seen, the F-statistic is directly related to the R^2 value. All else being equal, the higher the R^2 value, the higher the value of the F-statistic, and *vice versa*.

b) *The individual impact of a particular explanatory variable*

In this case the appropriate test statistic is the t-statistic (denoted by t_k). It is defined as:

³⁶ Strictly speaking, a test statistic follows the identified probability distribution under the assumption that the null hypothesis is valid.



$$t_k = \frac{b_k - \beta_k^*}{\text{se}(b_k)} = \frac{b_k}{\text{se}(b_k)} \quad \text{if } \beta_k^* = 0$$

Again, it is a random variable, which is distributed according to the t probability distribution, with $(n - K)$ degrees of freedom.

The t-statistic is inversely related to the standard error of the estimator. As such, the value of the statistic is affected by the sample size, the degree of variation in the sample values of X_k , and the degree of correlation between X_k and the other explanatory variables. In particular, if there is a high degree of correlation between some, or all, of the explanatory variables (a situation known as *multicollinearity*), then the value of the t-statistic may be very low.

F3.3 Determining the significance of the test statistic

Given the computed value of the test statistic, the final step is to decide whether this value is unlikely to have arisen under the associated probability distribution, and hence that the null hypothesis is unlikely to be true. That is, does the computed value the test statistic justify the rejection of the null hypothesis – at the chosen level of significance (α) ?

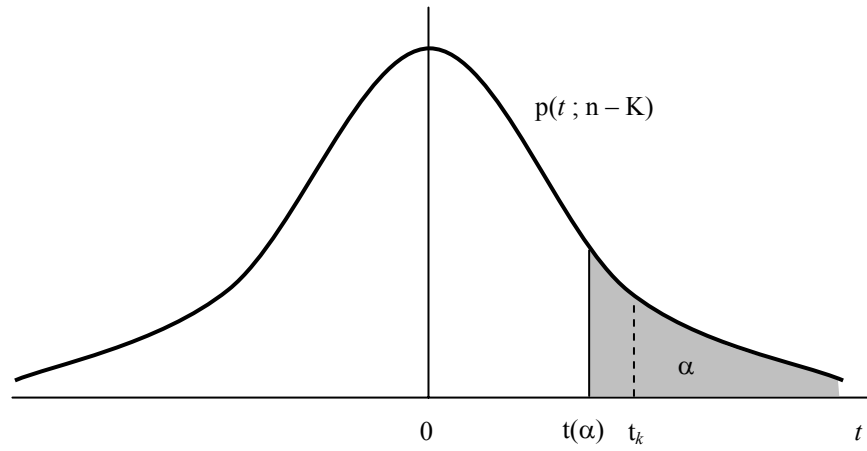
Essentially, there are two ways in which this is question can be answered. The first is to compare the computed value of the test statistic with a *critical value* (or critical values) that reflect(s) the level of significance that has been chosen. The second is to calculate a probability value (or *p-value*) for the test statistic, and then compare this directly with the chosen level of significance. The two alternative approaches are illustrated in the three panels of Figure A2 for the case of the t-statistic, where the bell-shaped curve is the probability distribution for the statistic with $n - K$ degrees of freedom (i.e. $p(t ; n - K)$). For any arbitrary value t^* , the area under the curve to the right of t^* represents the probability that the computed value of the test statistic is greater than this value when the null hypothesis is true. Similarly, the area under the curve to the left of t^* represents the probability that it is less than this value.

For a one-way hypothesis (i.e. panel a), the critical value $t(\alpha)$ is chosen so that the probability that the computed value of t_k is greater than $t(\alpha)$ – i.e. the grey shaded area – is equal to the chosen level of significance α . The test statistic is said to be *significant* if it is greater than this critical value – as is the case in Figure F2. In this case the null hypothesis is rejected, and one can conclude that $\beta_k > 0$. That is, the coefficient is significantly greater than zero. By construction, the probability of this conclusion being incorrect (i.e. a type I error) is equal to α .

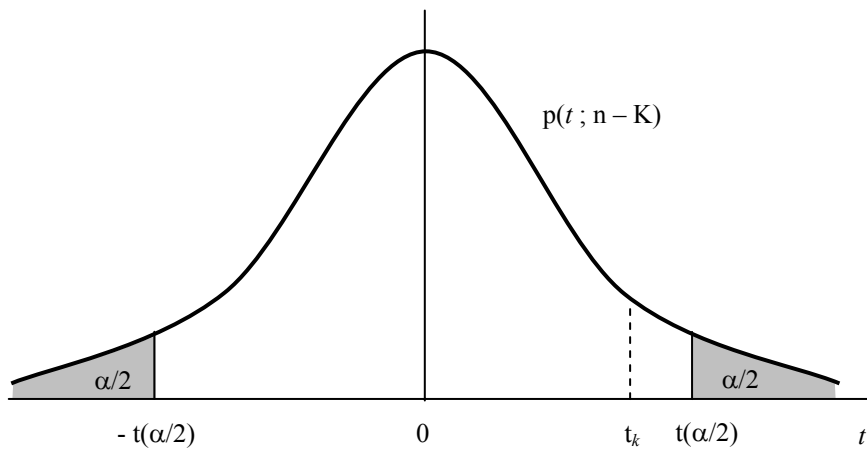
For a two-way hypothesis (i.e. panel b) there are two critical values – an upper value and a lower value. These values are chosen so that the probabilities of t_k being greater than the upper value, or of it being less than the lower value – i.e. the two grey shaded areas – are both equal to $\alpha/2$. Since the t distribution is symmetric, the magnitudes of the two critical values are the same. If t_k is greater than the upper value, or less than the lower value, then the null hypothesis is rejected, and one can conclude that $\beta_k \neq 0$. Again, by construction, the probability of this conclusion being incorrect (i.e. a type I error) is equal to α .

Figure F2: Test for individual coefficient b_k

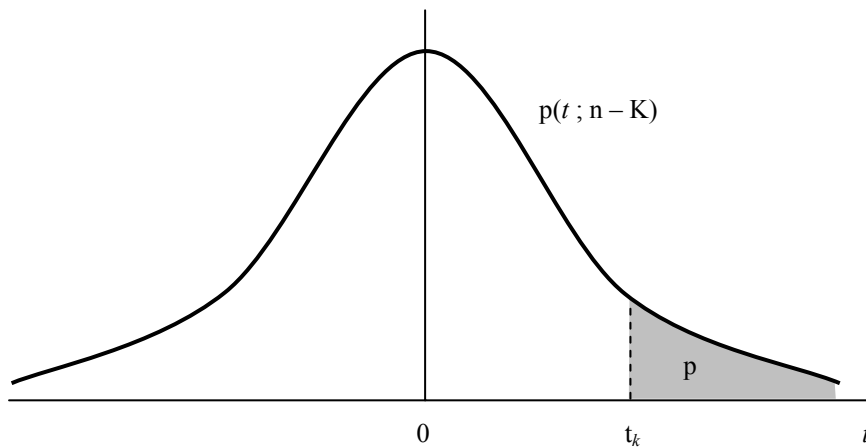
a) One-way hypothesis: level of significance = α



b) Two-way hypothesis: level of significance = α



b) One-way hypothesis: probability value = p





There are two important points to note from the first two panels of Figure F2. The first is that because $t(\alpha)$ is less than $t(\alpha/2)$, the computed test statistic can be significant under a one-way hypothesis, but not significant under a two-way hypothesis. That is, one may be able to conclude that $b_k > 0$, but not be able to conclude that $b_k \neq 0$! This anomalous situation highlights the importance of defining the hypothesis before one performs the test.

The second point is that if the level of significance is reduced continuously (i.e. the grey shaded areas shrink), then eventually any value for t_k will become insignificant, and hence the null hypothesis will be no longer be rejected. The value of α at which this occurs is called the probability value (or *p-value*) for the statistic. Panel c) shows the p-value for t_k under a one-way hypothesis. This represents the minimum level of significance for which the null hypothesis would be rejected. If the p-value is smaller than the value of α that has been chosen for the test – as is the case in Figure A2, then the null hypothesis is rejected.

The critical value approach and the p-value approach are, of course, just different perspectives on the same decision. However, the advantage of the p-value approach is that it gives an immediate indication of the “degree” of significance, or insignificance of the coefficient. For example, if the level of significance for the test has been set at 0.1, it is useful to know whether a coefficient has been deemed insignificant because its p-value is equal to 0.11, or because it is equal to 0.99. Because of this advantage, p-values have been shown for the individual coefficients in all of the results tables rather than the computed t-statistics.

While the shape of the probability F distribution differs from that of the t distribution shown in Figure A2, the procedures for determining the critical value and the p-value for the F-statistic are essentially the same as those described for the one-way hypothesis in panels a) and c).

F4. Interpretation of the measures

Ideally one would like to be in a situation where the R^2 / adjusted R^2 value is high; the F-statistic is significant; the t-statistics are significant for all of the model variables; and the coefficients exhibit the expected signs and magnitudes. In this case, the hypothesised PRF is unambiguously a good model (although not necessarily the best possible model), and one can identify the impacts of the individual explanatory variables.

Unfortunately, a more common outcome is that some measures are high / significant, while others are not. This can occur for a number of reasons. In particular, if there is a high degree of correlation between the values of the explanatory variables (i.e. high multicollinearity), then it is possible for the F-statistic to be significant, but for all the individual t-statistics to be insignificant. That is, one can conclude that the collectively the variables have an impact, but one cannot distinguish the impacts of the individual variables. This is more often a reflection of the “quality” of the sample data, than that of the model. Unless the variables are correlated for the population as a whole, the apparent anomaly may disappear if a new sample is taken, or the current sample enlarged.

Furthermore, it is perfectly possible to have a high R^2 value, but for the F-statistic to be insignificant; or to have a low R^2 value and a significant F-statistic. For example, if $n = 45$ and $K = 5$, the resultant F-statistic is significant at a 5% level of significance for all values



of R^2 greater than 0.21. In contrast, if $n = 21$ and $K = 11$ then it is insignificant for all values of R^2 less than 0.75.

In these situations, where the measures give apparently conflicting messages, the most important measure to consider is the F-statistic. If this is insignificant at the chosen level of significance, then it implies that the PRF is no better at explaining the variations in the dependent variable than a trivial model made up of a constant and a random term. Of course, the fact that the F-statistic is significant may be of little comfort if the objective of the analysis is to evaluate the individual impact of a particular variable.

While the F-statistic is the best measure for assessing the overall validity of a particular model, it is not generally meaningful to compare alternative models (with different sets of explanatory variables) on the basis of their respective F-statistic values. Unless the number of explanatory variables is the same in each model, one cannot conclude that the model with the higher value is the better model. There are a number of formal tests that can be used to compare alternative (non-nested) models. However, a simple way of doing so is to compare the adjusted R^2 values of the two models – assuming that the respective F-statistics are both significant. While there is no formal test that can be used to determine whether any difference is statistically significant, the model with the higher value may be considered preferable.