

CONFRONTING THEORY WITH EVIDENCE: METHODS &
APPLICATIONS

CONFRONTING THEORY WITH
EVIDENCE: METHODS & APPLICATIONS

A Thesis Submitted to the School of Graduate Studies in Partial
Fulfillment of the Requirements for the Degree of Doctor of
Philosophy

STEPHANIE M. THOMAS, B.A., M.A.

McMaster University ©Copyright by Stephanie Thomas, September 29, 2016

McMaster University DOCTOR OF PHILOSOPHY (2016) Hamilton, Ontario
(Economics)

TITLE: Confronting Theory with Evidence: Methods and Applications

AUTHOR: Stephanie M. Thomas, B.A. (McMaster University), M.A.
(University of Western Ontario)

SUPERVISOR: Professor Jeremiah Hurley

NUMBER OF PAGES: (xviii), 240

1 Abstract

Empirical economics frequently involves testing whether a theoretical proposition is evident in a data set. This thesis explores methods for confronting such theoretical propositions with evidence. Chapter 1 develops a methodological framework for assessing whether binary ('Yes'/'No') observations exhibit a discrete change, confronting a theoretical model with data from an experiment investigating the effect of introducing a private finance option into a public system of finance. Chapter 2 expands the framework to identify two discrete changes, applying the method to the evaluation of adherence to clinical practice guidelines. The framework uses a combination of existing analytical techniques and provides results which are robust and visually intuitive. The overall result is a methodology for evaluation of guideline adherence which leverages existing patient care records and is generalizable across clinical contexts. An application to a set of field data on supplemental oxygen administration decisions of volunteer medical first responders illustrates. Chapter 3 compares the results of two mechanisms used to control industrial emissions. Cap and Trade imposes an absolute cap on emissions and any emission capacity not utilized by a firm can be sold to other firms via tradable permits. In Intensity Targets systems firms earn (owe) tradable credits for emissions below (above) a baseline implied by a relative Intensity Target. Cap and Trade is commonly believed to be superior to Intensity Targets because the relative Intensity Target subsidizes emissions. Chapter 3 reports on an experiment designed to test theoretical predictions in a long-run laboratory environment in which firms make emission abatement technology and output production decisions when demand for output is uncertain, and banking of tradable permits may or may not be permitted. Particular focus is placed on testing whether the flexibility inherent to Intensity Targets can lead them to be superior to Cap and Trade when demand is stochastic.

2 Acknowledgments

Thank you to my supervisor, Jeremiah Hurley, for the countless edits, suggestions and support for this project, as well as for encouraging me to take my own path in research. I would also like to thank the members of my supervisory committee. Jeffrey S. Racine, for the amazingly prompt turnaround for all of my questions, regardless of his geographic location. Neil J. Buckley for the hours of collaboration through the development and implementation of our experiment, preparation for conferences and incredible academic support throughout my Ph.D. And to Stuart Mestelman, your attention to the details of my personal development as an academic has been pivotal to my success in the PhD program and completion of this dissertation. It has been fantastic working with you all.

I would like to acknowledge funding support from the Canadian Social Sciences and Humanities Research Council (Award #865-2008-0060), without which my research presented in Chapter 3 on emission trading markets would not be possible.

I would also like to thank the research team at the McEEL lab. This thesis would not exist without you. Our weekly meetings were a familiar constant that kept me going through rough times. Your unwavering confidence in me, and willingness to support new ideas and new approaches provided the ideal environment for me to develop confidence as an experimentalist and as an economist.

Special thank you also to Professor Bram Cadsby for inspiring the development of Chapter 1, and helpful comments from Professors Magee, Johri, and Sweetman and my peers in the graduate economics seminar. Thank you to Professor Feeny for insightful comments on Chapter 2.

Finally, my deepest gratitude to those who have supported me throughout

my Ph.D. My parents, who made sure I was well-housed and had excellent computing technology during every stage of my University studies, especially in this past year. And G.S. for restoring hope and passion to my world. There is room for even more now that this is complete.

Table of Contents

1	Abstract	iii
2	Acknowledgments	iv
3	List of Tables	x
4	List of Figures	xiv
5	List of Abbreviations	xvii
6	Declaration of Academic Achievement	xviii
	Introduction	1
	References	4
1	Playing by the rules? Agreement between predicted and observed binary choices	5
1.1	Introduction	5
1.2	Methodology	7
1.3	Theoretical predictions vs. observations	10
1.3.1	A theoretical model of participation	10
1.3.2	Observations	12
1.3.3	Matching theoretical predictions and observations	14
1.4	Observations	16
1.4.1	Empirical	18
1.4.2	Standard	20
1.4.2.1	Including interactions	24

1.4.3	Nonparametric	28
1.4.3.1	Bandwidth selection	30
1.4.4	Comparison of approaches	33
1.4.4.1	Confusion matrices and correct classification ratio	34
1.4.4.2	Receiver operator characteristics curves	36
1.4.4.3	Youden's J	38
1.4.4.4	Cohen's κ	40
1.5	Identifying switch-points	41
1.5.1	Identifying candidates using observations: The cumulative summation method	41
1.5.2	Identifying candidates using predicted values: Youden's optimal J	45
1.5.3	Identifying candidates using predicted values: maximum absolute gradients	47
1.5.4	Comparison of candidate switch-points	49
1.6	Results	50
1.6.1	Bootstrapping procedure	50
1.6.2	Observations vs theoretical predictions	51
1.7	Conclusion and discussion	54
	References	57
	Appendix 1.A Frequency of observations	59
	Appendix 1.B Standard approach	59
	Appendix 1.C Standard approach with interaction	60
	Appendix 1.D Coefficients of determination	60
	Appendix 1.E Bandwidth selection	61
	Appendix 1.F Is <i>income</i> relevant?	62
	Appendix 1.G Receiver operator characteristics curve	64

Appendix 1.H	Cumulative summation intersection	67
Appendix 1.I	Identifying candidates using observations: search method	67
Appendix 1.J	Gradients	69
Appendix 1.K	A strength measure for OYJ candidates	71
2	A Standardized Method for the Evaluation of Adherence to Practice Guidelines	74
2.1	Introduction	74
2.2	Data	79
2.2.1	Missing vital signs	83
2.2.2	Contact with emergency services	85
2.3	Method stage 1	87
2.3.1	Summary measures of correct classification	88
2.4	Application: Stage 1 adherence results	91
2.4.1	The economic consequences of non-adherence	94
2.5	Method stage 2	96
2.6	Application: Stage 2 adherence results	100
2.6.1	Relative performance of stage 2	103
2.6.1.1	Empirical strategy	105
2.6.1.2	Linear approach: regression with higher order terms	106
2.6.1.3	Search approach: the best fit guideline	109
2.6.1.4	Comparison	111
2.7	Conclusion and discussion	116
	References	119
Appendix 2.A	Observations without imputed missing values	123
Appendix 2.B	Nonparametric Estimation	125
Appendix 2.C	Nonparametric results with missing values	127
Appendix 2.D	Data sharing agreement	129

3 Cap and Trade versus Intensity Targets: Emissions trading markets with stochastic demand	131
3.1 Introduction	131
3.2 Literature	136
3.3 Experimental design	140
3.4 Theory	141
3.4.1 A dynamic stochastic model of emission permit trading with banking	142
3.4.2 A static model of emission permit trading	146
3.5 Experimental parameterization and procedures	151
3.5.1 Parameterization	152
3.5.2 Procedures	161
3.5.2.1 Efficiency	162
3.5.3 Summary of parameterization and procedures	163
3.6 Predictions and results	163
3.6.1 Results by treatment	168
3.6.2 Efficiency	170
3.6.3 Results by firm type	171
3.6.4 Emission permit prices	173
3.6.4.1 Emissions and permit price volatility	174
3.6.5 Intensity choices and risk attitudes	176
3.7 Discussion and conclusions	179
References	182
Appendix 3.A Supplementary Figures	186
Appendix 3.B Experiment instructions and screen captures	193
Discussion and conclusions	236
References	240

List of Tables

1.1	Summary of observations.	14
1.2	Classification matrix of observations and theoretical model predictions.	14
1.3	Probit regression estimates	21
1.4	Probit regression estimates with interaction terms	24
1.5	Bandwidths generated using least squares cross validation and maximum likelihood cross validation	31
1.6	Outline of a confusion matrix	34
1.7	Adjusted correct classification ratios for each approach.	35
1.8	Area under the receiver operator characteristics curve for each approach.	37
1.9	P-values of bootstrap tests of differences in areas under receiver operator characteristics curves for each approach.	38
1.10	Youden's J values for each approach.	39
1.11	Cohen's kappa values for each approach.	40
1.12	Candidate switch-points identified by intersection of cumulative summation and inverse cumulative summation of participate outcomes.	42
1.13	Switch-points identified using the optimal Youden's J value.	46
1.14	Switch-points identified using the maximum absolute gradient.	48

1.15	MAG and OYJ method switch-point candidates and confidence intervals for each approach.	51
1.A.1	Number of observations by income and contribution rate	59
1.F.1	Bandwidths generated using least squares cross validation (with and without income) and maximum likelihood cross validation.	62
1.I.1	Candidate switch-points by search approach and optimal Youden's J method.	68
1.I.2	Candidate switch-points by search approach and maximum correct proportion method.	68
1.I.3	Candidate switch-points by search approach and maximum Cohen's kappa method	69
1.J.1	Gradients at the MAG candidate switch-points.	70
1.J.2	Gradients at the OYJ candidate switch-points.	71
1.K.1	Strength measure of candidate switch-points identified using the optimal Youden's J method.	73
2.1	Variables used in analysis of supplemental oxygen decisions.	83
2.2	Classification matrix guide.	88
2.3	Classification matrix summary measures.	89
2.4	Classification of AUC values.	90
2.5	Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called.	92
2.6	Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called.	93
2.7	Classification matrix summary measures by EMS contact sub-group.	93
2.8	Annual cost of oxygen delivery by EMS contact sub-group.	95
2.9	Candidate steps and bootstrapped 90 percent confidence intervals using the nonparametric strategy.	103

2.10	Candidate steps of the empirical strategy and bootstrapped 90 percent confidence intervals.	105
2.11	Linear regression estimates with interaction and squared terms.	107
2.12	Candidate steps of the linear strategy and bootstrapped 90 percent confidence intervals.	109
2.13	Candidate steps of the search strategy and bootstrapped 90 percent confidence intervals.	110
2.14	AUC values of each approach	112
2.15	P-values of bootstrap tests of differences in areas under ROC curves.	113
2.A.1	Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called for data without imputed missing values.	124
2.A.2	Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called for data without imputed missing values.	124
2.A.3	Classification matrix summary measures by EMS contact sub-group for data without imputed missing values.	125
2.A.4	Annual cost of oxygen delivery by EMS contact sub-group for data without imputed missing values.	125
2.B.1	Bandwidths generated using least squares cross validation for data with and without imputed missing values.	127
2.C.1	Candidate steps of the nonparametric strategy for data without imputed missing values and bootstrapped 90 percent confidence intervals.	128
3.1	Parameters used to generate net marginal revenue schedules.	155
3.2	Net marginal revenue of clean firms.	155
3.3	Net marginal revenue of dirty firms.	155

3.4	Demand sequences.	157
3.5	Experiment parameters and predictions.	160
3.6	Summary of predictions and observed results by treatment pooled across firm types.	168
3.7	Summary of efficiency measure predictions and results by treat- ment pooled across firm types.	170
3.8	Summary of predictions and results by treatment and firm type.	172
3.9	Summary of mean permit prices by output demand state and treat- ment banking status, pooled across firm types.	174
3.10	Summary of mean variance and volatility by treatment, pooled across firm types.	175
3.11	Correlation of 'closeness' to target and level of risk aversion. . . .	177

List of Figures

1.1	Predictions of the theoretical model of participation by income and contribution rate.	12
1.2	Participation observations by contribution rate and income. . . .	18
1.3	Proportion of participation by contribution rate and income using the Empirical approach.	20
1.4	Predicted probability of participation by contribution rate and income using the Standard approach.	22
1.5	Predicted probability of participation by contribution rate and income using the Standard approach with interactions.	25
1.6	Predicted probability of participation by contribution rate and income using the Nonparametric approach.	32
1.7	Receiver operator characteristics curves for each approach.	36
1.8	Switch-point identification using the intersection of cumulative summation of 'Do not Participate' and inverse cumulative summation of 'Participate' outcomes for each approach.	43
1.9	Example of a perfect switch-point.	44
1.10	Example of absence of a switch-point.	44
1.11	Identification of switch-points by mapping optimal Youden's J values to contribution rates.	45
1.12	Gradients by contribution rate and income for each approach. . .	48

1.13	Candidate switch-points of the MAG and OYJ methods for the Non-parametric approach.	52
1.14	Candidate switch-points of the MAG and OYJ methods for the Standard approach.	53
1.15	Candidate switch-points of the MAG and OYJ methods for the Standard approach with interactions.	54
1.F.1	Predicted probability of participation by contribution rate (with and without income) using the Nonparametric approach.	64
2.1	Recommended and observed oxygen administration decisions by respiration rate.	81
2.2	Illustrated data summary.	85
2.3	Illustrated data summary by EMS contact sub-group.	87
2.4	Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach.	102
2.5	Predicted administration of oxygen using the Empirical approach by respiration rate and EMS contact sub-group.	106
2.6	Predicted administration of oxygen using the linear approach by respiration rate and EMS contact sub-group.	108
2.7	Optimal guideline using the search approach by respiration rate and EMS contact sub-group.	111
2.8	Predicted administration of oxygen for smooth approaches by respiration rate and EMS contact sub-group.	114
2.9	Candidate steps and confidence intervals by approach and EMS contact sub-group.	116
2.A.1	Illustrated data summary for data without imputed missing values.	123

2.A.2	Illustrated data summary by EMS contact sub-group for data without imputed missing values.	124
2.C.1	Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach for data without imputed missing values.	128
3.1	Implied output demand schedules.	154
3.2	Cap and Trade emission permit demand schedules: Output margin.	158
3.3	Cap and Trade marginal abatement cost curves: Intensity margin.	159
3.4	Correlation of 'closeness' to target and level of risk aversion. . . .	177
3.5	Correlation of average permits banked and level of risk aversion.	178
3.A.1	Box-plots of session results and predictions by treatment.	187
3.A.2	Box-plots of session efficiency results and predictions by treatment.	188
3.A.3	Box-plots of session results and predictions by treatment and firm type.	189
3.A.4	Box-plots of session emission permit price results and predictions by treatment and output demand state.	190
3.A.5	Box-plots of session volatility and variance results and predictions by treatment.	191
3.A.6	Average levels of risk aversion and deviation from the intensity target, by session.	192

5 List of Abbreviations

CCR	Correct classification ratio
adj-CCR	Adjusted Correct classification ratio
AUC	Area under the curve
ROC	Receiver operator characteristics curve
AIC	Akaike's information criterion
CS	Cumulative summation
ICS	Inverse cumulative summation
CSI	Cumulative summation and inverse cumulative summation intersection
CSM	Cumulative summation method
YOJ	Youden's optimal J
MAG	Maximum absolute gradient
TPR	True positive rate
TNR	True negative rate
FPR	False positive rate
TP	True positive
TN	True negative
FP	False positive
FN	False negative
MFR	Medical First Responder
PCR	Patient Care Report
EMS	Emergency Medical Services
SPM	Social Profit Maximization
NMR	Net Marginal Revenue
MAC	Marginal abatement cost

6 Declaration of Academic Achievement

The following is a declaration that the content of the research in this document has been completed by Stephanie Thomas. Dr. Neil Buckley, Dr. Jeremiah Hurley, Dr. Stuart Mestelman, and Dr. Jeffrey S. Racine provided insightful advice and manuscript review for all chapters of this work.

Chapters 1 and 2 are wholly the work of Stephanie Thomas. In Chapter 3 Stephanie Thomas contributed to the experimental design and programming of the experiment, was responsible for data collection and analysis as well as writing of the manuscript. Dr. Neil Buckley assisted with the experimental design as well as contributing funding for the experiment.

Introduction

Lionel Robbins (1932) described economics as ‘the science which studies human behavior as a relationship between given ends and scarce means which have alternative uses’. While this description appears straightforward, some debate regarding the scientific nature of economics persists. A New York Times opinion piece by Krugman (2013) entitled ‘Maybe Economics is a Science, But Many Economists Are Not Scientists,’ highlights the broad audience of the controversy. According to the Merriam-Webster (2016) dictionary science is ‘knowledge about, or study of, the natural world based on facts learned through experiments and observation,’ and according to the Science Council (2009) ‘the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.’ Both definitions hinge upon systematically seeking out and using evidence to understand the world. As the title suggests, this thesis investigates methods for using evidence (data) to evaluate theoretical propositions, thereby placing this economic work squarely within the domain of science. With an overarching theme of a scientific approach, this work seeks out evidence and takes a systematic approach to evaluating that evidence, specifically whether or not a particular theoretical proposition is apparent within a given data set. Each chapter incorporates evidence from a different source. Data from two economics experiments and one administrative data set are used. Evaluation of the evidence is carried out using a combination of meth-

ods and intuitive graphics. Two chapters develop an innovative and flexible framework for assessment of evidence against pre-existing theoretical propositions and guidelines. The third chapter evaluates the relative performance of two market mechanisms for the control of industrial emissions using data generated in an economics experiment. Overall, the aim is to understand whether or not the evidence support or reject pre-defined theoretical propositions.

Chapter 1 begins with an investigation of the match of a set of data gathered in the experimental economics laboratory to a theoretical structure hypothesized to account for behaviour. The main issue in this chapter is that the theoretical structure consists of a single discrete change in an outcome which consists of two levels and ranges over a continuous covariate. A flexible non-parametric conditional density approach is adopted and contrasted to a range of alternatives. The Nonparametric approach is shown to be particularly useful for illustrating relationships within a set of experimental lab data.

Chapter 2 builds upon the framework of Chapter 1, extending both the complexity of the theoretical structure (and therefore the analytical framework) and the potential applicability of the result. In this chapter an administrative data set consisting of the decisions of medical practitioners to administer supplemental oxygen to members of the community is compared to the practice guideline which should govern these decisions. Issues of data quality are addressed in a simple fashion and an economic analysis of the impact of perfect adherence to the guideline is undertaken. The key contribution of the chapter is the generalizable framework for the evaluation of clinical practice guidelines which is developed.

Lastly, Chapter 3 describes an economic experiment designed to answer the question of whether or not Cap and Trade emissions permit trading programs are the most effective choice, relative to Intensity Target systems, for

controlling pollution in a setting when demand for output is uncertain. This chapter focuses heavily on the careful generation of the experimental data, with analysis of the results being much more straightforward than in Chapters 1 and 2.

The results of this work explicitly offer value to three groups: practicing analysts of experimental economics laboratory data, assessors of health care system performance, and governing bodies tasked with developing markets for emission permits. However, the methods of gathering data through careful experimentation and the frameworks of analysis developed in this volume are not strictly limited to these groups. In fact, the approaches used here are broadly applicable to any scientist seeking an alternative approach to confronting a theoretical structure with evidence.

References

- Krugman, P. (2013). "Maybe Economics Is A Science, But Many Economists Are Not Scientists". In: *The New York Times, The Opinion Pages*. October 21.
- Merriam-Webster (2016). *Science*. URL: <http://www.merriam-webster.com/dictionary/science> (visited on 03/19/2016).
- Robbins, L. (1932). "The nature and significance of economic science". In: *The Philosophy of Economics: An Anthology*, pp. 73–99.
- Science Council (2009). *Our definition of science*. URL: <http://sciencecouncil.org/about-us/our-definition-of-science/> (visited on 03/19/2016).

Chapter 1

Playing by the rules? Agreement between predicted and observed binary choices

1.1 Introduction

The objective of this work is to determine whether data collected from economic activities support the predictions derived from a theoretical model of economic behaviour. The specific prediction described here is a step-wise relationship between a binary 'Yes'/'No' outcome variable, and two explanatory variables. For example, a consumer may decide to purchase a product over a range of low prices, so the outcome is 'Yes', and decide not to purchase at prices above a certain price threshold, so the outcome becomes 'No' from there on. In the raw form, the observations would include some noise due perhaps to impulse purchases or lack of attention. Using a standard parametric framework for smoothing such noise the location of a discrete change is masked, prompting the investigation of alternative approaches to evaluation. Using flexible nonparametric regression opens the possibility of locating a

discrete change or 'switch-point' within the smoothed observations. Three methods for locating candidate switch-points are suggested. The final proposed framework for evaluation uses the nonparametric smoothing approach combined with a maximum absolute gradient switch-point candidate identification strategy. The candidate switch-points are compared to the predictions of the theoretical model via constructing nonparametric bootstrapped confidence intervals which acknowledge the interdependence of the observations. The theoretical model and observations used to illustrate the techniques discussed here are taken from a single treatment of the experiment reported in Buckley et al. (2015). The observations consist of a set of observed decisions to participate or not participate in an activity and are conditioned on two explanatory variables. The methods developed here maintain as strict an independence as possible between the observations and the predictions of the theoretical model. This means that rather than attempting to explain the matches of the theory with the observations, the theory is defined first, and, if necessary, the observations are secondly smoothed using a regression framework. The first and second parts are then compared. This means that if the researcher wishes to compare a different theory to the observations it is straightforward to do so and requires no alteration to the description of the smoothed observations. It also means that if one suspects that the smoothed observations are improperly described that the regression framework can be altered independently of the theoretical predictions. An advantage of this approach is that the results can be intuitively illustrated, which offers substantial appeal to researchers wishing to communicate with diverse audiences. Applications of the methodological framework extend naturally to archival data and data from field experiments as well as to controlled laboratory experiments. Examples of these applications include decisions to look for a job or not (labor force participation) or adherence to professional practice guide-

lines in accounting, law or medicine (See Chapter 2 of this thesis).

1.2 Methodology

Empirical economics frequently involves testing whether the predictions of a theoretical model are realized under controlled conditions. This paper proposes a new method for assessing whether binary ('Yes'/'No') observations ranging over a continuous covariate exhibit a discrete change which is consistent with an underlying theoretical model. An application using observations from a controlled laboratory environment illustrates the method, however, the methodology can be used for testing for a discrete change in any binary outcome variable which occurs over a continuous covariate such as medical practice guidelines, firm entry and exit decisions, labour market decisions and many others. The observations are optimally smoothed using a Nonparametric approach which is demonstrated to be superior, judged by four common criteria for such settings. Next, using the smoothed observations, two novel methods for assessment of a step pattern are proposed. Finally, nonparametric bootstrapped confidence intervals are used to evaluate the match of the pattern of the observed responses to that predicted by the theoretical model. The key methodological contributions are three innovative methods proposed for assessing the step pattern. The promise of this approach is illustrated in an application to a controlled experimental lab data set, while the methods are easily extendable to many other settings.

Once the basic overview of the match with theory is established using a classification matrix approach the discussion moves on to smoothing, comparing three techniques to achieving this objective while incorporating the effects of two covariates. The first technique, an 'Empirical approach,' simply calculates basic proportions. The next is the standard parametric technique in which

a probit estimation strategy is employed.¹ The last is a Nonparametric approach in which the conditional density of the positive participation decisions is estimated. An Appendix compares the approaches in detail. The results of each smoothed model are then compared to the theoretical predictions and statistical significance established using a bootstrapped confidence interval approach.

A key weakness of the parametric approach is illustrated here. The parametric technique suggests very tight confidence intervals but is clearly misspecified. Using the Nonparametric approach the observations demonstrate reasonable support for the theoretical model. Evidence for a match of the candidate switch-points with those suggested by the theoretical model is found in most instances using the final proposed framework. Section 1.3 describes the theory and observations, Section 1.4 deals with smoothing the observations. Section 1.5 proposes two new methods for identifying switch-points, and Section 1.6 constructs nonparametric bootstrapped confidence intervals to evaluate the match of the candidate switch-points with the predictions of the theoretical model. Section 1.7 concludes, discussing alternatives and extensions of the framework.

The techniques explored here were developed in response to a particular situation arising in the experimental lab in which a theoretical model suggested that the outcome of interest would exhibit a clearly defined cutoff. There were few well defined options for accepting or rejecting the suggestions of the theoretical model, and the more standard options produced uninformative results. The insights gathered in the process of investigation and presented here offer a new direction for the confrontation of theory with evidence.

¹All results were also carried out using a logit technique with virtually indistinguishable results.

In addition to the analysis presented here variants of structural breaks and regression discontinuity were considered, treating the continuous covariate as the variable over which breaks occur, as year does in the familiar macroeconomic sense of structural breaks. In the case of testing for unknown structural breaks the single model of the relationship between the outcome and the two covariates proposes 5 breaks in the continuous covariate, each dependent upon the ordered covariate level. Testing for every possible combination of breaks implies 118,755 tests.² Taking the approach of a known break one could also simply suppose the true breaks to be those of the theoretical model and test for a match. In the macroeconomic context, this approach suggests that the model takes on a different form before and after the break. In the case of a binary outcome variable this implies a model for the positive outcomes of *participate* and another for the negative outcomes. Conceptually this would imply a belief that the covariates have different influences upon the participation decision based on whether the particular participation decision is taken before or after a certain level in the continuous covariate. Similarly, regression discontinuity suggests that the application of a treatment effect at a known break has the potential to result in two different models before and after the break. In this experiment participants chose the outcome variable based on the two covariates where the continuous covariate was determined in the experiment and not applied as a treatment variable. As an adaptation, the match of the observations against all possible cutoffs is provided in the paper, using the goodness of fit metrics of Correct Classification Ratio, Adjusted Correct Classification Ratio, The Area Under the Receiver Operator Characteristics Curve and Cohen's κ . Despite being able to select a maximum value in order to identify a candidate switch point, the true behavior of the observations is masked by this technique because a break in the sense of a

²The number of ways to choose 5 break-points from 29 possibilities

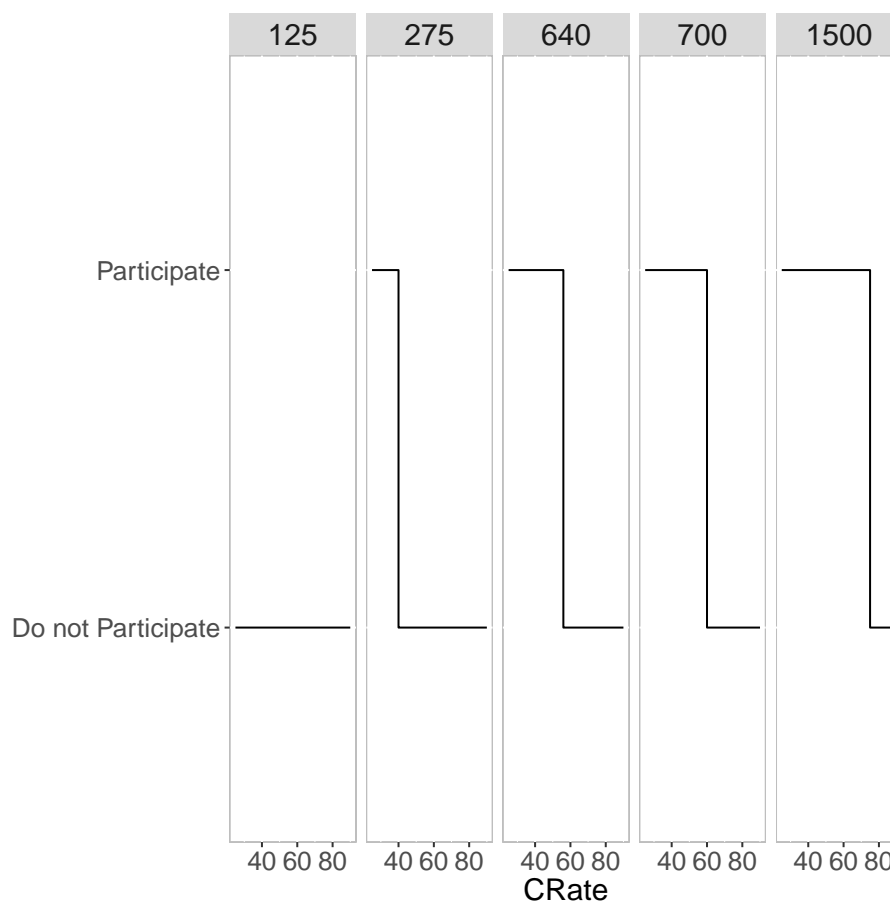
clear shift in the observations may or may not actually exist.

1.3 Theoretical predictions vs. observations

1.3.1 A theoretical model of participation

The theory and observations forming the basis of the example illustrated in this paper were drawn from an experiment presented in Buckley et al. (2015). The experiment sought to explore the role of mixed finance arrangements whereby a private good is funded publicly but also available to purchase privately. Participants in the experiment were asked to submit a preferred contribution rate to a fund financing public provision of the good, and then contributed to the public fund at a rate determined by the median of the preferred contribution rates. This rate is the tax rate that is applied to the incomes of all participants in the session and provides resources to publicly provide the private good. Any income remaining could be used to purchase additional amounts of the private good in a manner defined by the particular treatment. The public fund was invested and each participant then received an equal share of the total fund, i.e. a private good. For the purpose of this investigation only the theory and observations of the ‘top-up’ treatment in Buckley et al. (2015) are used. The focus here is on the development of a method for determining whether or not behavior within a treatment supports a theoretical proposition. In this context, participation, referred to as ‘*participate*’ in this paper, is defined as ‘topping-up’ in Buckley et al. (2015); that is, purchasing an amount of the good privately in addition to the level provided through the public system. Non-participation is defined as consuming only the amount provided publicly. The decision to participate is dependent upon ‘*income*’ and the tax rate (labelled here as the contribution rate ‘*CRate*’).

The theoretical model of top up behaviour outlined in Buckley et al. (2015) describes a pattern of participation decisions which are dependent upon income and the rate at which all the participants in a group contributed to the public fund. For each level of income in the model there is a contribution rate at which participation switches from being an optimal to a non-optimal decision. The theoretical predictions thus follow, for each level of income, a pattern of participation in privately topping up the consumption of the publicly provided private good at low CRates and then stopping this topping up once the CRate reaches a sufficiently high value. This creates a distinct 'step' or 'switch' in the relation between topping up and not topping up as the CRate rises. The step patterns in Figure 1.1 illustrate the relationship. As income level increases the contribution rate at which participation becomes non-optimal increases.



Source: Buckley et al. (2015).

Figure 1.1: Predictions of the theoretical model of participation by income and contribution rate.

1.3.2 Observations

The set of observations consists of 500 participation decisions along with the associated income levels and contribution rates. In each period subjects first learned their income level (*income*) and then submitted their preferred rate of contribution to the public fund. The median of the submitted rates was selected as the *CRate* for the group and the *CRate* proportion of income deducted from each participant's funds. Each member of the group was free to purchase additional investments independently of the group fund with the remaining proportion of their income as part of the 'top-up' treatment.

If a participant made any additional purchases the variable '*participate*' was classed 'Participate', and as 'Do not Participate' otherwise.³

The data set includes decisions made by 50 individuals for each of 10 decision periods. The experiment was designed as a repeated one-shot game and no statistically significant period effects were reported in Buckley et al. (2015). Voting to determine the contribution rate to the public fund took place in 5-person groups. No statistically significant group effects were identified by Buckley et al. (2015). An independent observation of a determination of the *CRate* is defined as the result of a 5 person group in each period of the experiment, so the experiment contains 100 independent observations.

The outcome variable *participate* is an unordered factor variable taking on a value of 'Participate' if the individual purchases a nonzero amount of the good privately, and 'Do not Participate' otherwise. *Income* is a discrete ordered variable taking on values {125, 275, 640, 700, 1500} which were randomly assigned to individuals and were distributed so as to ensure that no two group members experienced the same level of income within a period. Each participant experienced each level of income twice during the 10 decision periods but was not informed of which level of income would occur prior to the start of any period and so experienced the assignment of income in a random manner. *CRate* is a continuous variable which could take on any value in the range [0, 100]. In the data set we observe 29 unique values in the range [25, 90]. Table 1.1 provides a summary of the data. The total number of observations is 500. 229 participant decisions involved purchasing a positive amount of private investment, 271 did not participate in private purchasing. There are 100 observations at each level of income. Contribution rates to the public fund ('*CRate*') range from 25 to 90, with mean 55.3 and median 54.5.

³The sum of the contributions to the public fund was invested and the fund increased in a pre-defined manner with $\frac{1}{5}$ of the total returned to each participant. Private purchases were also augmented in the same way but were not shared among group members.

See Appendix Section 1.A Table 1.A.1 for the frequency of observations of *CRate*.

participate	n	Income	n	CRate	value
		125	100	Minimum	25
Participate	229	275	100	Maximum	90
Do not Participate	271	640	100	Mean	55.3
		700	100	Median	54.5
		1500	100		

Table 1.1: Summary of observations.

1.3.3 Matching theoretical predictions and observations

		Observation		
		Do not Participate	Participate	Total
Theoretical Prediction	Do not Participate	223	54	277
	Participate	48	175	223
	Total	271	229	500

Table 1.2: Classification matrix of observations and theoretical model predictions.

A basic way to summarize the overall match of the collected observations with the predictions of the theoretical model can be done simply by classifying each observation based on whether or not the observation is in agreement with the appropriate theoretical prediction. There are four possible classes.

Class 1: the theoretical model predicts participation and participation was observed,

Class 2: the theoretical model predicts no participation and participation was not observed,

Class 3: the theoretical model predicts participation and participation was not observed ('under participation'), and

Class 4: the theoretical model predicts no participation and participation was observed ('over participation').

Table 1.2 presents the number of observations in each of the described classes in tabular form, commonly referred to as a 'classification matrix' or 'confusion matrix'. In this case 48 of a total 500 observations qualify as under-participation (10%) while 54 observations qualify as over-participation (11%). To evaluate the fit of the observations with the predictions of the theoretical model, the Correct Classification Ratio (CCR) is frequently used as a measure of accuracy. This measure is simply the proportion of observations which match the predictions of the theoretical model (i.e. Class 1 and Class 2). In this case, this refers to the accuracy of the theoretical model at producing predictions which describe the observations collected in the experiment. Here 80% of observations are in agreement with the predictions of the theoretical model. The CCR can be biased, however, because even if chosen by chance, there is a higher probability of simply choosing the most frequent outcome and being correct. Adjusting for the probability of choosing the most frequent outcome by chance this value falls to 55% using the adjusted correct classification ratio (adj-CCR, defined in Section 1.4.4). Another measure used to describe agreement between the observations and predictions is the the area under the receiver operator characteristics curve (AUC), which is described in detail in Section 1.4.4.2. AUC values range from 0.5 to 1, with higher values indicating greater levels of agreement between observations and predictions. In this case the value of the AUC is 0.79, which suggests substantial agreement between the observations and theoretical predictions, and is nearly identical to the CCR result. Finally, Cohen's κ describes the amount of agreement between the theoretical predictions and the observations beyond that occurring by chance, where a value of 1 indicates perfect agreement and 0 no agreement (Cohen, 1960). The value of Cohen's κ in this case is 59%, which

is similar to the adj-CCR result and serves to confirm substantial agreement beyond random chance between the theoretical predictions and the observations.

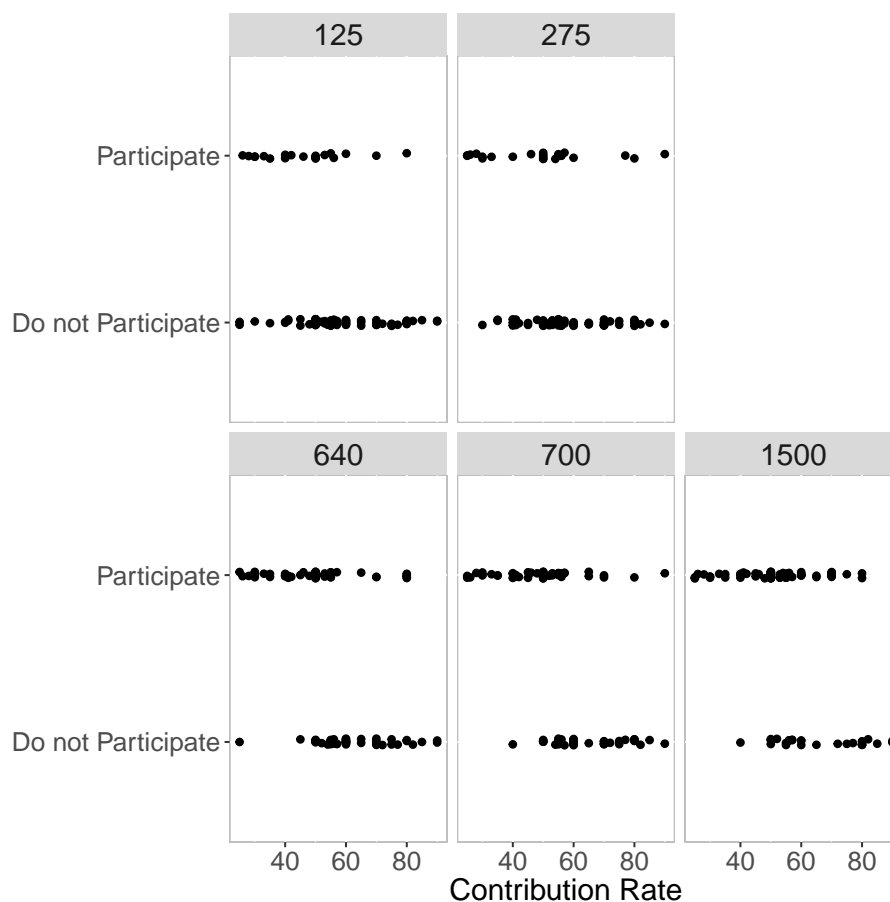
Of the subset of predictions which are in disagreement, 53% are cases which qualify as 'over participation' and 47% qualify as 'under participation'. A one sample proportions test with continuity correction fails to reject a null of equal proportions (p-value 0.6205), indicating that there is no reason to suspect any systematic tendency towards over- or under- participation among decisions which are not in agreement with the theoretical model.

The classification matrix technique is a simple means to assess the overall fit of the observations with the predictions of the theoretical model, but it does not address the relative influence of the explanatory variables upon the outcome or whether the observations do in fact exhibit a discrete 'step'. The remainder of this chapter will explore the relationship between *participate*, *income* and *CRate* in greater detail.

1.4 Observations: Describing the relationship between *participate*, *income* and *CRate*

This section presents three approaches to investigating the relationship between *participate*, *income* and *CRate*. The first method, the 'Empirical' approach, simply examines the frequency of participation at each combination of *income* and *CRate*. The second method, the 'Standard' approach, attempts to estimate a line of best fit using Probit regression. The third approach, the 'Nonparametric', estimates the relationship by nonparametric conditional density estimation which calculates an optimal bandwidth and uses kernel regression. Goodness of fit is assessed using four commonly used criteria: the adjusted correct classification ratio (adj-CCR), Cohen's κ (Cohen, 1960), area under the

receiver operator characteristics curve (AUC) and Youden's J (Youden, 1950). Each criteria is described in detail in Section 1.4.4. Of the smoothing strategies considered only the Nonparametric approach is capable of unmasking switch-points in the data. The Nonparametric approach also dominates the Standard approach in terms of fitting the observations on all four criteria. The purpose of investigating the relationship of *participate* with *income* and *CRate* in the observations is ultimately to assess whether the pattern of observations is similar to the step pattern suggested by the predictions of the theoretical model. A conclusive step pattern would present as two distinct groups of observations which do not overlap across *CRate* for each level of *income* and all that would be required would be to determine the *CRate* at which the observations of *participate* 'switch' from 'Participate' to 'Do not Participate'. Figure 1.2 presents the observations of the experiment and offers motivation for smoothing. In order to facilitate a visual assessment the observations have been jittered vertically in order to show multiple observations which occur at the same *CRate*. Each level of *income* is presented in a different pane. Despite these two visual adjustments it is clear that it is not possible to identify a clear 'switch' from 'Participate' to 'Do not Participate' for each level of *income*. The main objective of this section is to condense the observations into a form conducive to identifying a candidate switch-point, should it exist.



Source Buckley et al. (2015). Slight vertical jittering (displacement) of points to show multiple observations.

Figure 1.2: Participation observations by contribution rate and income.

1.4.1 The Empirical Approach

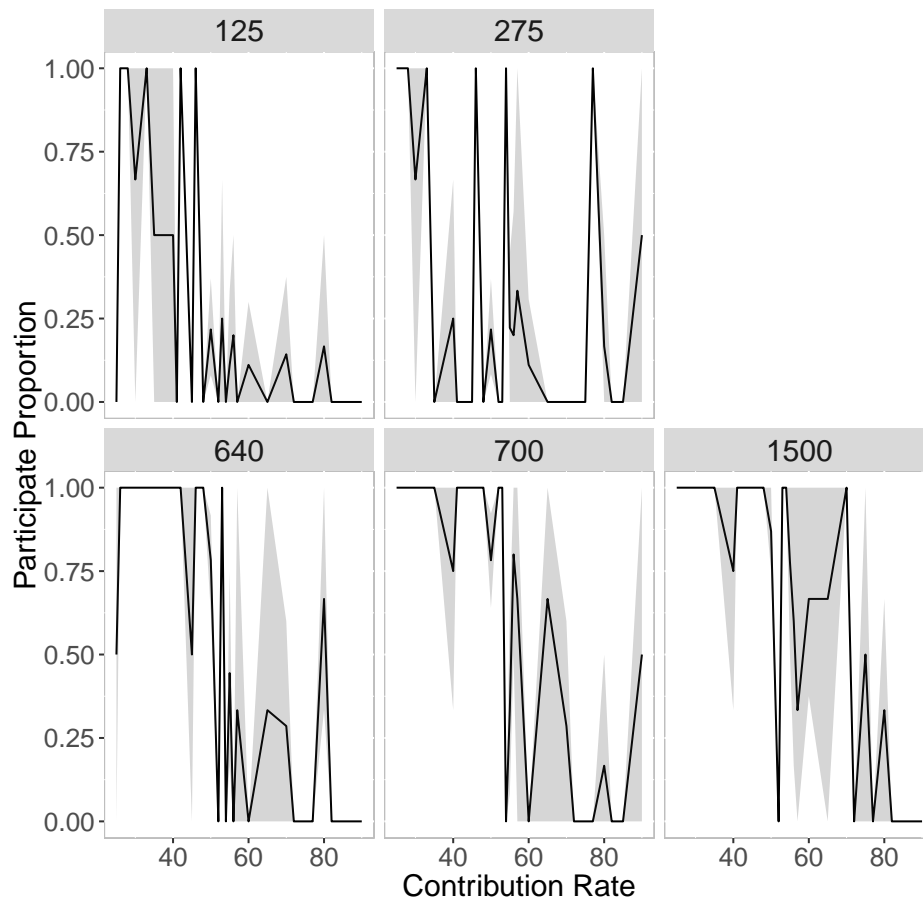
A Pearson Chi-squared test is used as a preliminary assessment of the existence a relationship between *participate*, *income* and *CRate*. This test proposes a null hypothesis of independence among all the three variables; failure to reject this null indicates a lack of a relationship. Here the null is rejected (p-value of 0.0000) suggesting that a relationship in fact exists.

The Pearson Chi-squared test is a frequency-based test which relies on comparing the observed frequencies with the expected frequencies if the null were true and no relationship were to exist. In this case *income* has 5 levels

and *CRate* is treated as a discrete variable of 29 values, so there are 145 frequencies to compare. In order to avoid a biased result the number of observations in each *income* - *CRate* cell should be greater than 5 as per the specifications of the Chi-square test. Table 1.A.1 in the Appendix Section 1.A shows that this is not the case for the observations at hand; many of the cells contain only one observation. As a rudimentary remedy to achieve the necessary minimum of 5 observations in each cell the Chi-squared test was conducted over a grouped *CRate* which was arbitrarily categorized into classes with values of less than or equal to cutoffs of $\{30, 40, 50, 60, 70, 80, 90\}$.⁴ The result is a rejection of the null of independence among the variables (p-value of 0.0000, as reported above). The test using the uncategorized *CRate* provided the same result (p-value of 0.0000). Both results are suggestive of the existence of a relationship.

The first method for condensing the raw observations is done by simply plotting the proportion of '*participate*' decisions which were to positively 'Participate' at each level of *income* and *CRate*. The noisiness of the results is indicative of the sensitivity of this approach to the number of observations at each *income*-*CRate* cell; nonetheless this is useful for gaining an initial idea of the relationships and suggests that the observations may in fact exhibit 'switch' type patterns. The lines in Figure 1.3 trace the proportion of positive 'Participate' decisions and suggest that, apart from noise, *participate* may exhibit distinct changes over *CRate* for at least some *income* levels. For instance, in the 125 pane the probability of *participate* falls sharply at a *CRate* of approximately 50. Similar changes are also visually identifiable in the 640 and 700 panes. The next section will attempt to address this noise using a Standard regression approach.

⁴The optimal bandwidth for 'CRate' is 3.1305, however this bandwidth results in less than the necessary 5 observations in each cell required to conduct the Chi-squared test. The arbitrary cutoff used here therefore represents an 'over-smoothed' comparison case.



Solid line is the proportion of 'Participate' outcomes. Grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 1.3: Proportion of participation by contribution rate and income using the Empirical approach.

1.4.2 The Standard approach

The main weakness of the Empirical approach is that the results are so noisy that multiple 'switches' in *participate* could potentially be identified for each *income* level. To smooth out this noise, the Standard approach described here employs Probit regression. Probit regression is among the most frequently used regression frameworks for estimating binary outcomes and so serves as a reference point for a broad audience of analysts. In this instance the approach treats *CRate* as a continuous variable, overcoming the pitfall of having too few observations in each *income-CRate* cell, and estimates a line of best

fit which is much smoother than that of the Empirical approach. The reduced noise should assist in identifying a unique candidate switch-point. This regression strategy is appropriate to the task of estimating proportions and can be alternatively interpreted as estimating the predicted probability of the decision *participate* being 'Participate' ⁵. The drawback of this approach, as will be shown, is that the results are so smooth that switch-points are masked completely.

The conditional probability $\widehat{\text{participate}}$ is defined here by:

$$\Pr(Y = 1|X) = \Phi(X'\hat{\beta}), \quad (1.1)$$

where in this case X is composed of the two explanatory variable vectors $X = (X_1 = \text{income}, X_2 = \text{CRate})$ and Y is the binary outcome variable *participate* which is conditional on X . $X'\hat{\beta}$ is referred to as the index function and the results are estimated by maximum likelihood estimation. Φ is the standard normal cumulative density function and is used to ensure that the predicted probabilities lie within the range $[0, 1]$. The complete details are provided in Appendix Section 1.B.

	Coefficient	Std. Error	z value	Pr(> z)
(Intercept)	1.44	0.30	4.82	0.00
CRate	-0.04	0.01	-8.41	0.00
Income275	0.11	0.21	0.52	0.61
Income640	1.03	0.20	5.05	0.00
Income700	1.18	0.20	5.76	0.00
Income1500	1.69	0.21	7.87	0.00

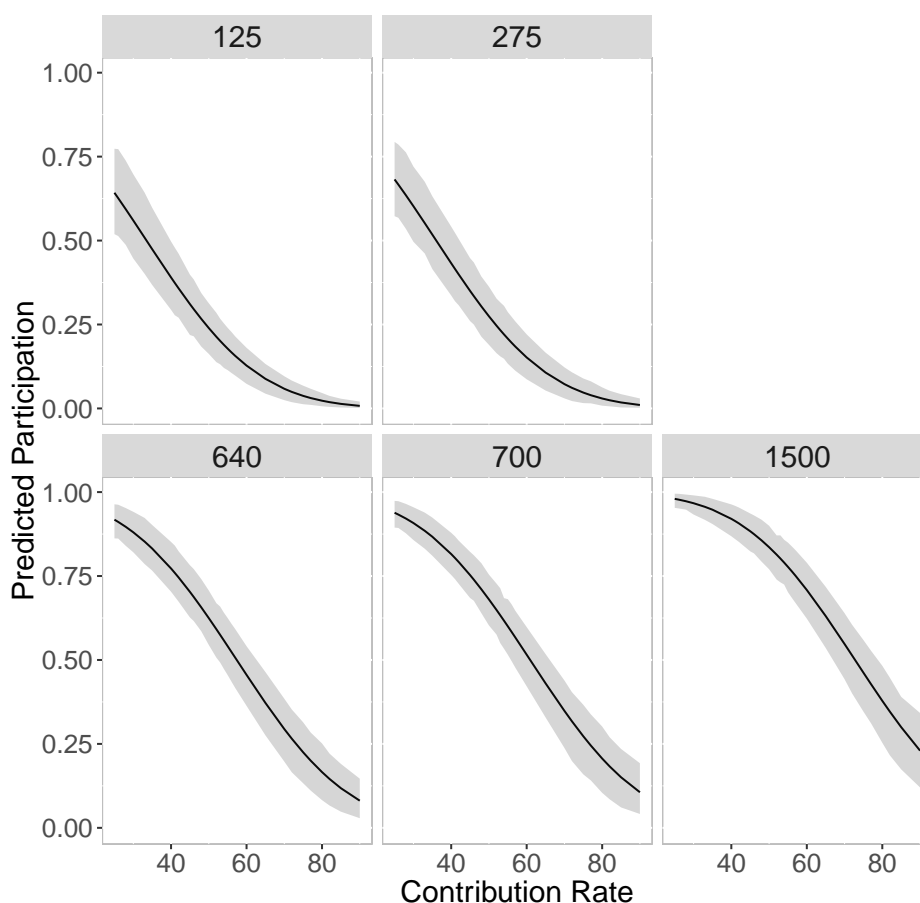
Table 1.3: Probit regression estimates

Table 1.3 provides the regression estimates ⁶. The outcome exhibits significant differences from the reference income level of 125 when participants face the

⁵Logistic regression for the odds of 'Participate' returned virtually identical results.

⁶The estimation is done using R's glm function in the stats package R Core Team (2015a), or the mfx package Fernihough (2014) which also provides marginal effects.

income levels of 640, 700 and 1500 but not the 275 income level. Both both the sign and ordering of the magnitudes are as expected. An income of 1500 has a larger effect on the estimated probability of participation than the 700 income level. The 700 income level, in turn, has a larger effect on the estimated probability than the 640 income level, and so on. The variable *CRate* also has the expected sign and is significant. As *CRate* increases the estimated probability of participation decreases.



The solid line is the probability of the 'Participate' outcome. The grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 1.4: Predicted probability of participation by contribution rate and income using the Standard approach.

The estimated probability of participation, $\widehat{participate}$, is illustrated for each *income* level in Figure 1.4, along with the bootstrapped confidence intervals

(discussed in more detail in Section 1.6). This figure demonstrates the exceptional smoothing achieved by the Standard approach. The multiple potential switches identified by the Empirical approach are now completely masked; no distinct switches are observable. Taking the results from Table 1.3 and Figure 1.4 together it is reasonable to infer that the probability of *participate* is declining as *CRate* increases and increasing as *income* increases.

For the probit model the predicted probabilities follow a smooth pattern by design, as dictated by the parametric structure. This is of no consequence if the population from which the data are drawn in fact follow this exact distribution. If our sample of observations is, however, not drawn from a population specified precisely by the parametric form estimated in Equation 1.1 then any inference derived from this model is misleading. Based on this fact one might be concerned with whether the probit model is a reasonable approximation to the population from which the data are drawn. A detailed examination of (pseudo) coefficients of determination which attempt to measure the amount of variation in the outcome (*participate*) attributable to variation in the explanatory variables *income*, *CRate*, is provided in Appendix Section 1.D. As well, Wald tests for the joint significance of all variables, income variables alone, and contribution rate alone are all rejected confirming that the variables are jointly significant. According to these commonly used metrics this model appears to deliver a fairly good fit to the observations, however, it does not suggest the discrete changes predicted by the theoretical model. Nor do any of these common tests target misspecification directly. Running a variant of Ramsey's RESET Test (Ramsey, 1969) suggested by Ramalho and Ramalho (2012) on the function described by:

$$Pr(Y = 1|X, \hat{y}^2) = \Phi(X'\beta + \theta\hat{y}^2) \quad (1.2)$$

reveals that the square of the fitted values of the original regression is significant (i.e. θ has a value of 5.43 and a p-value 2.4×10^{-4}) indicating that the null of correct specification should be rejected in favour of the alternative: that the model is misspecified. This invalidates any inference based upon this model because it is likely biased and inconsistent. In addition, confidence intervals are not proper confidence intervals since they are centered on a biased estimate.

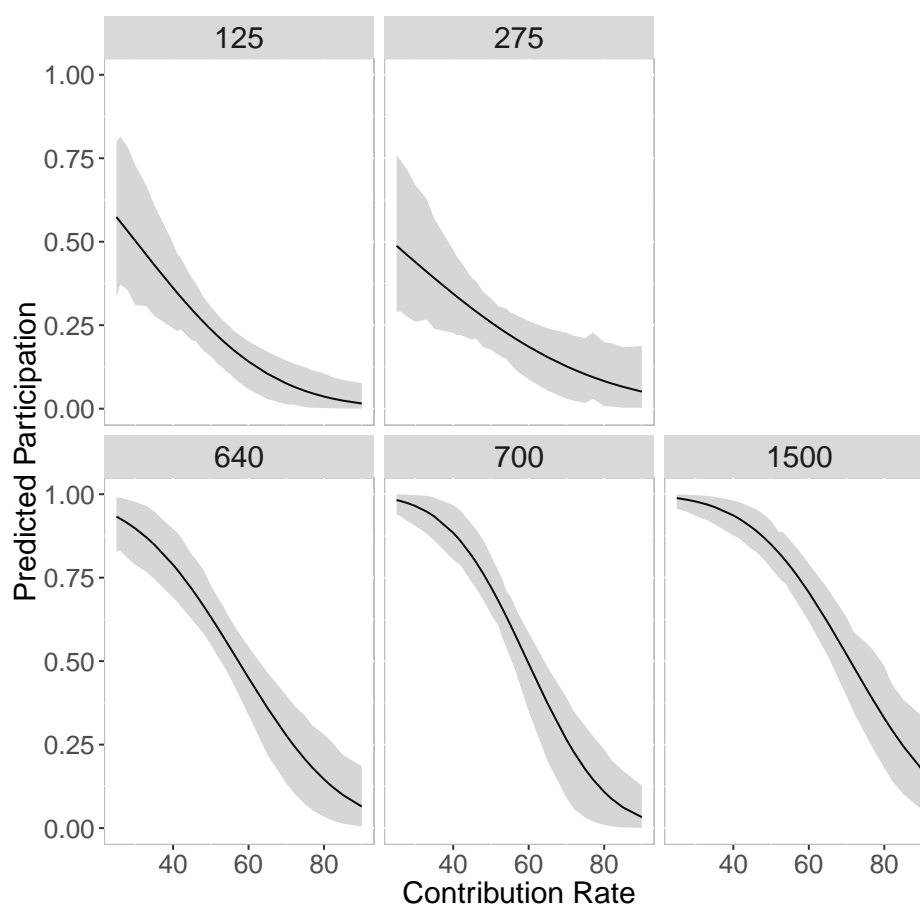
1.4.2.1 Including interactions

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.0883	0.6282	1.73	0.0832
CRate	-0.0361	0.0120	-3.01	0.0026
Income275	-0.5020	0.8540	-0.59	0.5566
Income640	1.5701	0.8858	1.77	0.0763
Income700	2.5412	0.9612	2.64	0.0082
Income1500	2.4063	0.9482	2.54	0.0112
CRate:Income275	0.0114	0.0160	0.71	0.4781
CRate:Income640	-0.0104	0.0163	-0.64	0.5247
CRate:Income700	-0.0248	0.0175	-1.42	0.1562
CRate:Income1500	-0.0132	0.0168	-0.78	0.4325

Table 1.4: Probit regression estimates with interaction terms

The RESET test run in the previous section suggests that the probit model is misspecified. While there is no way of knowing the particular form of the misspecification suggested by the RESET result, one possibility is that the model failed to account for potential interactive effects of *income* and *CRate*. Including this interaction changes the index function but does not improve the situation. Table 1.4 shows the results with the addition of an interaction between *income* and *CRate*. Again, in order to interpret the coefficients in Table 1.4 as probabilities the index function is distributed according to the cumulative normal distribution function. The results are similar to those without interactions, none of the interactive terms are significant. Reading the co-

efficients, including interactions, does not assist in describing the relationship. Figure 1.5 illustrates the predictions of the Standard model with interactions. Here the results still fail to clearly delineate a switch-point, a reflection the particular choice of model. The larger confidence bounds suggest that the inclusion of interactions leads to a loss of precision, however, both forms of the parametric approach fail to reject misspecification, which suggests that the results are inconsistent in both cases.



The solid line is the probability of the 'Participate' outcome. The grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 1.5: Predicted probability of participation by contribution rate and income using the Standard approach with interactions.

Comparing the results of the approaches without and with interactions, McFadden's Adjusted R^2 value (McFadden, 1973) is 0.24 for the model without

interactions and 0.23 with interactions, indicating that the inclusion of interactions does not assist in explaining variation in the decision to participate. The R^2 values of Cragg and Uhler (1970) are also commonly used to compare approaches and tell a similar story. The values are 0.4 for the model without interactions and 0.41 with interactions included, indicating that the approach with interactions does not offer much improvement in explanatory power over the model without interactions.

McFadden's and Cragg and Uhler's R^2 values are applicable only to the parametric approaches, whereas the Adjusted Count R_{AC}^2 value has the advantage that it does not depend on the approach used and so can be directly compared regardless of the approach to estimation. This metric is essentially the same as the adj-CCR described in Section 1.4.4, except that in this case the values of $\widehat{participate}$ are compared to the observed values of $participate$ by classifying the $\widehat{participate}$ with values greater or equal to 0.5 as positive 'Participate' decisions and all others as negative 'Do Not Participate' decisions. The statistic therefore summarizes the degree of match between the classified predictions and the observations. The value of this statistic is 0.5 without the interactions and 0.52 with interactions included, indicating a better fit in the case of the approach with interactions. Overall the difference between the approaches with and without interactions is small. When taken in conjunction with the lack of significance of the interaction terms, the overall results are supportive of concluding that the interaction terms are uninformative.

Another means of comparing results across models is via Akaike's Information Criterion (AIC) (Akaike, 1974) which describes the amount of information lost by using a model to describe a set of data. For the model with interactions the value is 526.19 which is larger than that of the probit estimation without the interaction term (524.36), indicating that including the interaction, while offering a worse fit according to McFadden's Adjusted R^2 , a better

fit according to Cragg and Uhler's R^2 , and a better fit according to the Adjusted Count R^2_{AC} results in a greater loss of information. The more parsimonious form, without interactions is preferable, indicating that the interaction terms possibly introduce a degree of multicollinearity. In support of the notion that the model is misspecified, the Ramsey RESET test variant again rejects the null of correct specification.⁷

A Wald test for the joint insignificance of the interaction coefficients confirms that these terms are not relevant in determining the participation decision, while the remaining coefficients remain jointly significant. Including the interaction term suggests about as good a fit to our data as the specification without interactions. The higher R^2 and AIC values combined with the rejection of the Ramsey RESET test variant suggest that the effect of *CRate* and/or *income* upon the *participate* decision is potentially more nonlinear than the specified probit model. Adding higher order terms in addition to, or instead of, the interaction term could improve the fit of this model. However, if we begin to adjust our model in order to achieve better results we run the risk of forcing the data to tell us the story we want to hear.

The Standard approaches illustrated here produce smooth declines in the predicted values of $\widehat{participate}$. These smooth declines suggest rejection of a hypothesis of switching patterns in the observations. Yet, the smoothness is largely the result of choosing to employ the probit technique, which, as was shown, is not a correct specification of the relationship between the variables at hand. In the next section an alternative smoothing technique will be explored, and in Section 1.4.4 a comparison of all the approaches is presented.

⁷The square of the predicted values in the regression (as in equation 1.2) is significant (p-value 0.01365)

1.4.3 The Nonparametric approach

The Standard approach investigated in the previous section smoothed the observations but did so in such a way as to completely mask the switch patterns observed under the unsmoothed Empirical approach. This section considers a Nonparametric alternative which smooths noise effectively while revealing switch patterns in the observations. Nonparametric regression presents a robust alternative to the Standard parametric approaches; along with being insensitive to a small proportion of outliers in the observations, these methods circumvent issues of model misspecification and have excellent in-sample fit. Because a specific form is not specified at the outset of the investigation the estimates of $\widehat{participate}$ may take on any shape, including those of the Standard approach. For experimentalists seeking to investigate relationships within relatively small, but carefully collected data sets these features are particularly attractive and simple to implement.

The Nonparametric approach relies upon the data at hand to form predictions, smoothing weighted observations within small sections of data called bandwidths. First an optimal bandwidth for each variable is calculated by minimizing a cross validation function. Then, the observations within each bandwidth are weighted according to a specified weight function and combined to produce a product kernel. This approach automatically takes into account any interactions and has the ability to exclude variables which are not relevant. By circumventing the need to choose the form of an estimating model this approach avoids issues of misspecification while retaining the ability to reproduce the results of any Standard approach. The nonparametric alternative of conditional density estimation as described first by Stone (1977) and more recently by Hall, Racine, and Li (2004) is implemented here in R using the np package developed by Hayfield and Racine (2008).

The problem at hand is to estimate the conditional density function $g(y|x) = \frac{f(x,y)}{\mu(x)}$ where $f(x,y)$ is the joint probability distribution of the outcome y and explanatory variables x and $\mu(x)$ is the mean of explanatory variables x . This is done via estimating the function:

$$\hat{g}(y|x) = \frac{\hat{f}(x,y)}{\hat{\mu}(x)} \quad (1.3)$$

Under the approach described by Li and Racine (2007) the numerator and denominator of the conditional probability function are described by:

$$\hat{f}(x,y) = \frac{1}{n} \sum_{i=1}^n K_{\gamma}(x, X_i) k_{\lambda_0}(y, Y_i) \quad (1.4)$$

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n K_{\gamma}(x, X_i), \quad (1.5)$$

with $K_{\gamma}(x, X_i)$ and $k_{\lambda_0}(y, Y_i)$ representing kernel density functions.

In this study the unordered dependent variable *participate* is estimated using the kernel suggested by Aitchison and Aitken (1976) and defined by:

$$\begin{aligned} k_{\lambda_0}(y, Y_i) &= l(Y_{is}, y_s, \lambda_s) \\ &= \begin{cases} 1 - \lambda_s & \text{if } Y_{is} = y_s \\ \frac{\lambda_s}{c_s - 1} & \text{if } Y_{is} \neq y_s \end{cases}, \end{aligned} \quad (1.6)$$

where y_s can take on c_s ordered values $0, 1, c_s - 1$. If $\lambda_s = 0$ then $l(Y_{is}, y_s, \lambda_s) = 1$ is an indicator function, and if $\lambda_s = \frac{c_s - 1}{c_s}$, then $l(Y_{is}, y_s, \frac{c_s - 1}{c_s}) = \frac{1}{c_s}$, a constant. Thus the range for the smoothing parameter associated with *participate* is $[0, \frac{2-1}{2} = 0.5]$.

The dependent variables both enter into the product kernel $K_{\gamma}(x, X_i)$ which

takes the general form:

$$K_\gamma(x, X_i) = W_h(x^c, X_i^c)L(x^d, X_i^d, \lambda), \quad (1.7)$$

where $\gamma = (h, \lambda)$ is a vector of continuous and discrete bandwidths, in this case for *CRate* and *income*. The superscript c denotes the continuous variable *CRate* and d the discrete variable *income*. The ordered levels of *income* are estimated using the kernel proposed by Racine and Li (2004) while the Epanechnikov (1969) kernel is used for the continuous variable *CRate*.

The Racine and Li (2004) kernel is described by:

$$L(x_i^d, x^d, \lambda) = \begin{cases} 1 & \text{if } |x_i^d - x| = 0, \\ \lambda^{x_i^d - x} & \text{if } |x_i^d - x| \geq 1 \end{cases}, \quad (1.8)$$

where λ must lie between 0 and 1.

The Epanechnikov (1969) kernel is defined by:

$$W(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2) & \text{if } u^2 < 5 \\ 0 & \text{otherwise} \end{cases} \quad (1.9)$$

$$\text{where } u = \frac{x^c - X_i^c}{h}$$

$$\text{and } h > 0,$$

1.4.3.1 Bandwidth selection

The choice of the particular kernel weight functions has little influence on the results of the nonparametric method while the bandwidth selection method has great impact (Li and Racine, 2007). Two bandwidth selection routines are considered here without altering the chosen kernels in order to investigate the impact of bandwidth selection upon the resulting estimates. Least

squares cross validation is the preferred method because it has the ability to remove irrelevant regressors ⁸ but it can be computationally intensive, and even prohibitive for large data sets ⁹. In this case, the routine takes less than one minute. An alternative to least squares cross validation is maximum likelihood cross validation, which can be less computationally intensive but has the drawback that it can oversmooth if the tails of the distribution are fat. This oversmoothing can potentially lead to an inconsistent estimate but it can also be beneficial if it effectively removes irrelevant regressors. The details of each of the methods are included in Appendix Section 1.E.

Variable	Least Squares	Maximum Likelihood	Upper Bound
Participate	0.0000	0.0512	0.5
Income	0.9926	0.9923	1
CRate	3.1305	3.1305	inf

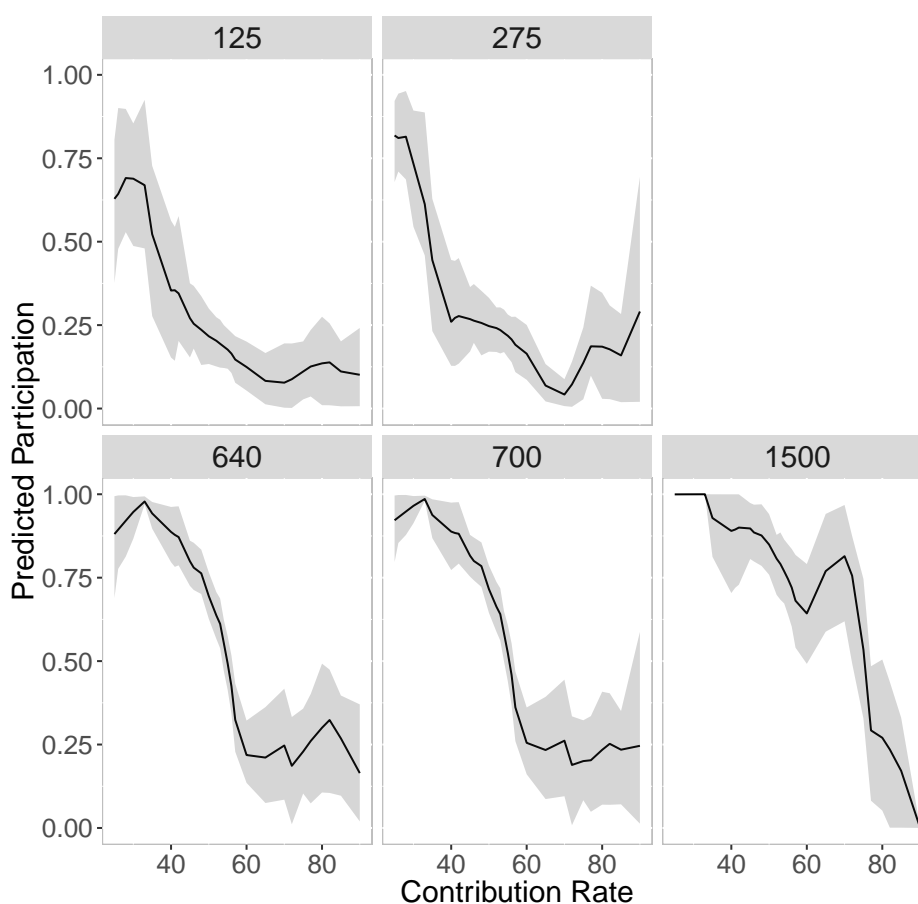
Table 1.5: Bandwidths generated using least squares cross validation and maximum likelihood cross validation

The bandwidths resulting from each selection method as well as the maximum value these bandwidths can take on given the chosen kernels are presented in Table 1.5. Bandwidths which are closer to their maximum values are indicative of variables which are irrelevant. The table suggests that *participate* and *CRate* are relevant, while *income* is very close to being smoothed out of the regression. In this case maximum likelihood cross validation is not suspected of oversmoothing, since the results of the least squares and maximum likelihood cross validation routines are very similar. Appendix Section 1.F investigates the relevance of *income* in detail, finding *income* to be highly relevant. In what follows, the least squares cross validated bandwidths of the approach with both *income* and *CRate* will be used.

⁸An irrelevant regressor is a variable whose variations do not contribute to variation in the outcome.

⁹Prohibitive computational intensity means that the routine may take months to determine a result using current computational technology.

Figure 1.6 illustrates the results of the Nonparametric approach for the predicted probability that *participate* = 'Participate': $\hat{g}(y|x) = \widehat{participate}$. The predictions retain some noise, but unlike both the completely unsmoothed Empirical approach or the oversmoothed Standard approach, these results support the existence of unique switch-points at each *income* level. While the confidence bounds are wider in some regions due to a lack of observations, the changes in $\widehat{participate}$ are much steeper than under the Standard approach. Section 1.4.4 will demonstrate the superiority of the Nonparametric approach to smoothing in terms of fitting with the observations.



Proportion of 'Participate' outcomes is the solid line. Grey shaded area is the bootstrapped 90 percent confidence interval.

Figure 1.6: Predicted probability of participation by contribution rate and income using the Nonparametric approach.

1.4.4 Comparison of approaches

The Nonparametric approach dominates in terms of producing predicted probabilities of participation which match with the collected observations on participation (in-sample fit). This better in-sample fit of the Nonparametric approach is a particularly attractive feature for analysts of experimental data because often the point of analysis is not to approximate an unknown population from which the results are a sample but to explore patterns within a carefully collected population of laboratory observations. As long as the number of explanatory variables are few, the relatively small data sets encountered by experimentalists are suitable for the least squares cross validation procedure which removes irrelevant variables with little to no risk of oversmoothing. In what follows, the use of the Nonparametric approach will serve as the reference case to evaluate the prediction that individuals will participate or will not participate in 'topping-up' consumption given the explanatory variables *income* and *CRate*. Comparisons will be made with the Standard approach as a point of illustration.

In order to compare approaches four criteria are presented here. The Adjusted Correct Classification Ratio (adj-CCR), or adjusted accuracy rate, forms the basic criterion for comparison (Fawcett, 2006). Two further measures include Receiver Operator Characteristics (ROC) curves (See Swets (2014) and Green and Swets (1966)) and Youden's J . As well, Cohen's κ , which adjusts for the probability of selecting the matching outcomes by chance, is provided. These approaches are considered because the traditional pseudo- R^2 values such as the Adjusted McFadden's R^2 reported for the results of the Standard approach are not comparable across models. In what follows, the Nonparametric approach shows the strongest performance regardless of the assessment criterion.

1.4.4.1 Confusion matrices and correct classification ratio

		Observed Participation	
		Do not Participate	Participate
Predicted Participation	Do not Participate	true negative	false negative
	Participate	false positive	true positive

Table 1.6: Outline of a confusion matrix

For each smoothing approach, a confusion matrix (or correct classification matrix) summarizes how well the resulting smoothed values match the originally observed outcomes. In Section 1.3 a classification matrix was introduced to compare the observations with the predictions of the theoretical model. Now, the same technique will be used to compare the observations with the predicted probabilities of participation for each of the Standard and Nonparametric approaches. For these comparisons, a 'threshold' must additionally be specified for sorting the predictions in to the classes 'Participate' and 'Do not Participate'. For example, if the Standard approach estimates that the conditional probability of participation is 70% for a participant with an *income* of 125 and a *CRate* of 55, then given a threshold such as 50% (the typical default threshold) this prediction would be classed as 'Participate' and compared to the actual observation. Whenever this positive decision matches the decision recorded in the experiment the total in the cell 'True Positive' increases. Table 1.6 provides the naming convention. If the prediction is positive but the actual observation was negative, a false positive is recorded whereas if the prediction was negative but the actual observation was positive a false negative is recorded.

The CCR is a simple way to summarize the entries in the confusion matrix. This measure consists of the fraction of outcomes which match the actual outcomes. As with the confusion matrix this measure is also dependent upon the particular threshold employed when converting the predicted probabil-

ities into binary outcomes. The default threshold used here for demonstration is 0.5. Unfortunately the CCR does not control for the probability of correctly choosing the more frequent outcome. To account for this, the adj-CCR is used. This measure is defined as:

$$\text{adj-CCR} = \frac{\text{True Positive} + \text{True Negative} - M}{n - M} \quad (1.10)$$

where M is the count of the most frequent outcome.

Approach	Threshold	Adjusted CCR	Lower Bound	Upper Bound
Empirical	0.5	0.66	0.55	0.75
Standard	0.5	0.50	0.41	0.62
Standard (Interaction)	0.5	0.52	0.41	0.63
Nonparametric	0.5	0.56	0.45	0.66

Values rounded to the nearest hundredth.

Bounds are bootstrapped 95 percent confidence intervals.

Table 1.7: Adjusted correct classification ratios for each approach.

The adj-CCRs for the approaches explored thus far are reported in Table 1.7. The Empirical approach offers the best fit and the Standard the worst. This is no surprise since the Empirical approach is completely responsive to the data at hand, however, this approach does not smooth noise very well, masking switch-type patterns in noise. The Standard approaches with and without interaction terms are very similar, adding interactions improves the fit by only 4%. As discussed earlier, while the Standard approach smooths out noise, it does so at the cost of masking potential switch-type patterns. The Nonparametric approach using least squares cross-validation offers a 12% improvement over the Standard approach, (8% over the Standard approach with interaction included). The Nonparametric approach is 15% worse than the Empirical approach, however, it clearly indicates a step pattern while outperforming

the Standard approaches.¹⁰

1.4.4.2 Receiver operator characteristics curves

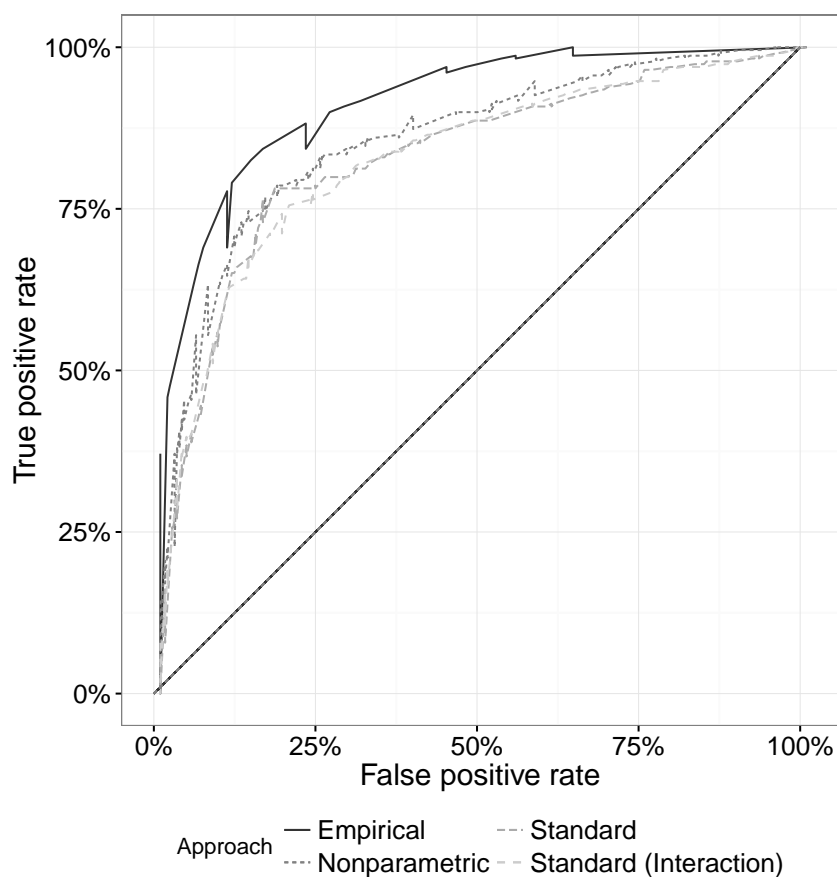


Figure 1.7: Receiver operator characteristics curves for each approach.

All the results of Table 1.7 are dependent upon the arbitrary choice of a threshold of 0.5. Receiver Operator Characteristic (ROC) curves, however, explore the effect of varying the threshold value, plotting the True Positive and False Positive Rates as the threshold is varied. These curves, presented in Figure 1.7, compare the predictive performance of the different estimation approaches

¹⁰For reference, the Nonparametric approach excluding *income* as a regressor (not shown in Table 1.7) worsens the fit by 16% over the Nonparametric approach including *income*, offering further support for including *income* in spite of the large smoothing parameter.

as the threshold changes. A detailed explanation of the Receiver Operator Characteristics Curve is provided in Appendix Section 1.G. More bowed out ROC curves indicate better predictive ability, so the Empirical approach is the best performing estimation approach and the two Standard approaches the poorest.

Area Under the ROC (AUC) is the preferred statistic used to quantify the ability of the results of a smoothing approach to fit an observed data set and for comparing ROC Curves. Perfect predictive ability produces an AUC of 100% while zero predictive ability produces an AUC of 50%. Table 1.8 presents these results and their bootstrapped 95% confidence intervals.¹¹ The Nonparametric dominates the smoothed approaches. The results indicate that the Empirical approach explains the data most effectively, while the Nonparametric approach is more efficient than the Standard approach.

Approach	AUC	Lower Bound	Upper Bound
Empirical	91.76	89.41	93.94
Standard	83.41	79.71	87.00
Standard (Interaction)	83.12	79.65	86.84
Nonparametric	86.15	82.68	89.47

Bounds are bootstrapped 95 percent confidence intervals.

Table 1.8: Area under the receiver operator characteristics curve for each approach.

The AUC corresponding to the Standard approach in Figure 1.7 is 83.41% and the AUC corresponding to the Standard technique with an interaction term is 83.12%, slightly lower than the original Standard approach. A bootstrap test of the difference between the two areas fails to reject the null of no difference.¹² This suggests that adding the interaction term to the Standard approach did not significantly improve the predictive ability of this approach.

¹¹Using the bootstrapping embedded within the Robin et al. (2011) package.

¹²Using the bootstrapping test provided within the Robin et al. (2011) package.

Approach	Standard (Interaction)	Nonparametric	Empirical
Standard	0.91057	0.27482	0.00
Standard (Interaction)		0.2215	0.00
Nonparametric			0.01

Table 1.9: P-values of bootstrap tests of differences in areas under receiver operator characteristics curves for each approach.

Additionally, testing the difference between the AUCs suggests that the advantage of the Nonparametric approach over the Standard approach is not significant. The p-value results using the bootstrapping test of the unpaired difference in AUCs provided within the pROC package of Robin et al. (2011) are presented in Table 1.9. Values less than 0.05 suggest rejection of the null hypothesis of no difference at the 5 percent level of significance. Only the Empirical approach is suggestive of a significant difference. While the improvement in predictive ability of the Nonparametric approach is small, the reduction in misspecification error combined with the substantial difference in capacity to visually suggest a switch-point afforded by this approach are important. This lends support to a descending ranking of preferability of the estimation strategies in terms of AUCs from least to most smoothed.

1.4.4.3 Youden's J

Another means for exploring the impact of varying the threshold upon the predictive power of the smoothing strategies is to compare Youden's J values (Youden, 1950) at each threshold. This index is described by:

$$J = \text{True Positive Rate} + \text{True Negative Rate} - 1, \quad (1.11)$$

where the True Positive Rate (TPR) and True Negative Rates (TNR) are defined by

$$TPR = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (1.12)$$

$$TNR = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \quad (1.13)$$

J is a measure of relative effectiveness and takes on a value of 0 if the true positive and true negative results are correctly classified at the same rate, and 1 if there are no false positives and no false negatives. Higher values of J are indicative of a more effective approach. The results from a threshold of 0.5 for each smoothing strategy are presented in Table 1.10. Larger values indicate better match of sorted predictions with the observations. Although the overlap in the confidence bounds of the Youden's J values indicates a lack of difference in the statistical sense, the relative ranking of the values is consistent with the adj-CCR and AUC results reported in Tables 1.7 and 1.8, respectively.

Approach	Threshold	J	Lower Bound	Upper Bound
Empirical	0.5	0.69	0.59	0.77
Standard	0.5	0.56	0.45	0.65
Standard (Interaction)	0.5	0.56	0.45	0.66
Nonparametric	0.5	0.59	0.49	0.68

Threshold for sorting predictions is 0.5.

Bounds are bootstrapped 95 percent confidence intervals.

Table 1.10: Youden's J values for each approach.

This measure again depends on the particular threshold employed. In Appendix Section 1.I a method which searches for a threshold by maximizing Youden's J is investigated.

1.4.4.4 Cohen's κ

Cohen's κ is a measure of the amount of agreement between the predictions and the observations beyond that occurring by chance. The measure is defined as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (1.14)$$

$$\text{where } p_0 = \frac{\text{true positive} + \text{true negative}}{\text{Total}}$$

$$\text{and } p_e = \frac{(\text{true negative} + \text{false negative})}{\text{Total}} * \frac{(\text{true negative} + \text{false positive})}{\text{Total}} + \frac{(\text{false positive} + \text{true positive})}{\text{Total}} * \frac{(\text{false negative} + \text{true positive})}{\text{Total}}$$

Approach	Threshold	kappa	Lower Bound	Upper Bound
Empirical	0.5	0.69	0.62	0.74
Standard	0.5	0.54	0.47	0.61
Standard (Interaction)	0.5	0.56	0.48	0.63
Nonparametric	0.5	0.59	0.52	0.66

Threshold for sorting predictions is 0.5.

Bounds are bootstrapped 95 percent confidence intervals.

Table 1.11: Cohen's kappa values for each approach.

A kappa value of 1 indicates perfect agreement and 0 no agreement. Table 1.11 presents the results of sorting the predictions of each approach according to the arbitrary threshold value of 0.5 and calculating Cohen's κ . The results indicate that again, the Empirical approach offers the most agreement between predictions and observations, but among the smoothing options the Nonparametric approach outperforms the parametric approach. The ranking is supported by the results of bootstrapped 95% confidence intervals. The overlap of the confidence intervals of the Empirical and Nonparametric approaches suggests a lack of a statistically significant difference. However, the Empirical approach is significantly different from the Standard approach suggesting that the Nonparametric approach offers a relevant compromise

between the over-smoothed Standard and under-smoothed Empirical approaches. In Appendix Section 1.I a method which searches for a threshold by maximizing Cohen's κ is investigated.

1.5 Identifying switch-points

1.5.1 Identifying candidates using observations: The cumulative summation method

Within a step-type pattern the 'switch-point' is the x axis coordinate marking the location of the step. With the observations at hand, this could be thought of simply as the point at which the 'Do not Participate' outcome becomes relatively more frequent than the 'Participate' outcome. A way to locate the *CRate* where this occurs is to simply plot the cumulative summation (CS) of the 'Do not Participate' decisions against the inverse of the cumulative summation (ICS) of the 'Participate' decisions on the same graph and locate the *CRate* where the two lines intersect; as presented in Figure 1.8. This intersection is called the Cumulative Summation Intersection (CSI) and the method used to identify the CSI is the Cumulative Summation Method (CSM). The details are contained in Appendix Section 1.H.

Figure 1.8 presents the CS and ICS data by *CRate* using solid and dotted lines respectively. The contribution rate where the distributions cross is taken as the estimated switch-point and is represented by the dashed lines in the figure. While one attractive feature of this method is that it does not require any smoothing, it applies equally well to smoothed predictions as long as these are classified according a threshold (as was done when calculating the CCR). Taking a 0.5 threshold for classification of the predictions, Table 1.12 presents the candidate switch-points for the Empirical, Standard and Nonparametric

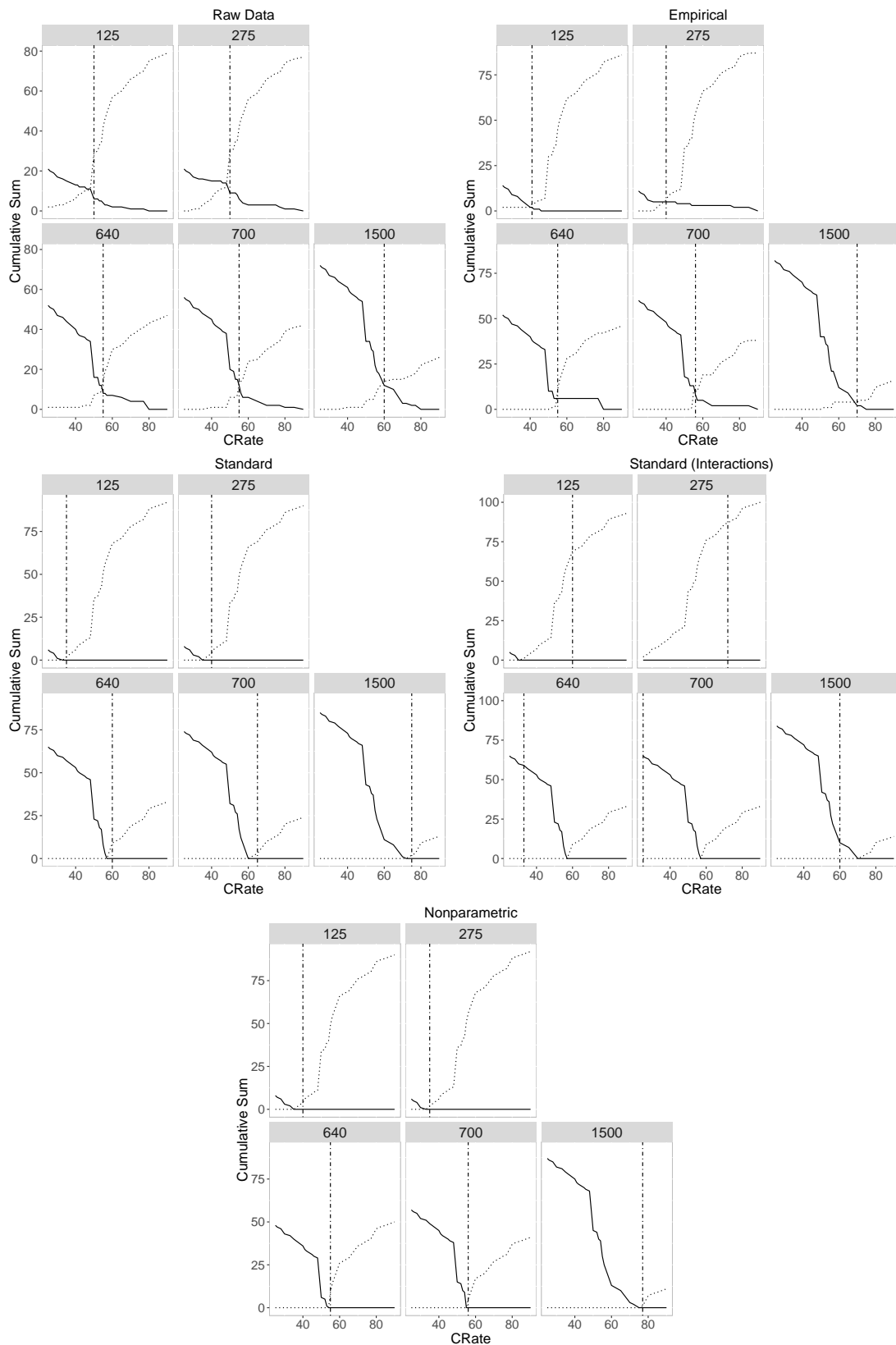
approaches, as well as directly to the raw data. The first $CRate$ at which the CS exceeds the ICS is the CSI and taken as the switch-point.

Switch-points identified directly with the raw data are convenient because no estimation is required, however this technique cannot differentiate perfect switch-points from equally distributed data with no switch-point. To illustrate this point, the CSI diagrams of two example data sets are presented. The first data set features data which are distributed in such a way as to represent a perfect switch. The second example incorporates data which are distributed equally and therefore have no switch at all. Both illustrative data sets are composed of 0's and 1's associated with a range from 1 to 100 and both with a mean of 0.5. Figures 1.9 and 1.10 show the weakness of applying the CSM to the raw data. In Figure 1.9 the data are presented in the first pane and have a clear switch-point. In the second pane, the CS and ICS are plotted and suggest a switch-point at 0.5, an exact match to the obvious location of the switch. Figure 1.10 however presents a data set with no switch-point at all in the first pane. The CSI obtained by plotting the CS and ICS suggests a switch point at 0.5. This method thus loses validity as the data become less representative of a switch. One might consider a method for incorporating the strength of the switch by recognizing that the perfect switch occurs at a cumulative sum of 50 in the case with a switch and at 25 in the non-switch case, but this is not explored here.

Income	Data	Empirical	Standard	Standard (Interaction)	Nonparametric
125	50	41	35	33	40
275	50	40	40	25	35
640	55	55	60	60	55
700	55	56	65	60	56
1500	60	70	75	72	77

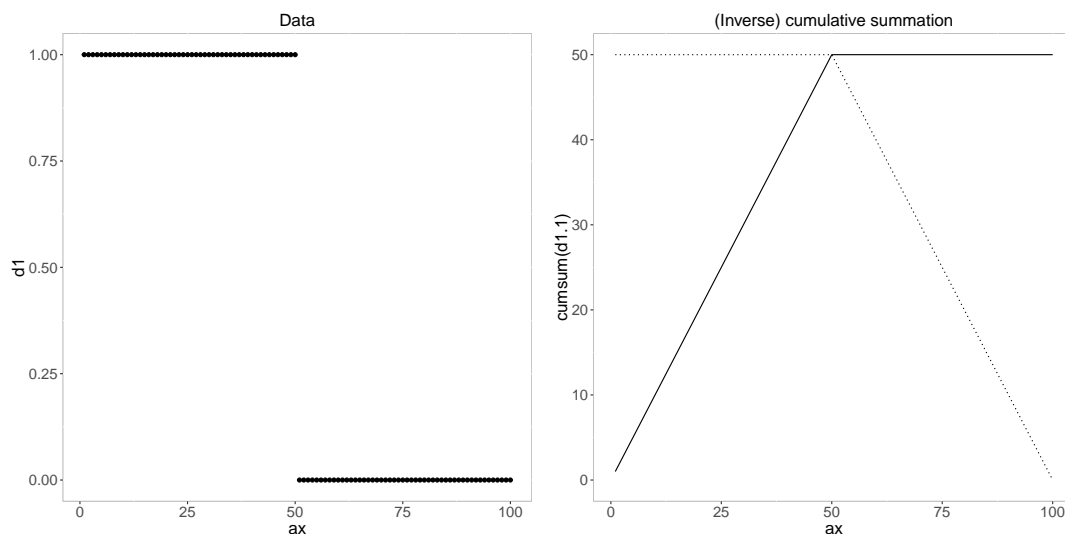
Predictions classified using a threshold of 0.5.

Table 1.12: Candidate switch-points identified by intersection of cumulative summation and inverse cumulative summation of participate outcomes.



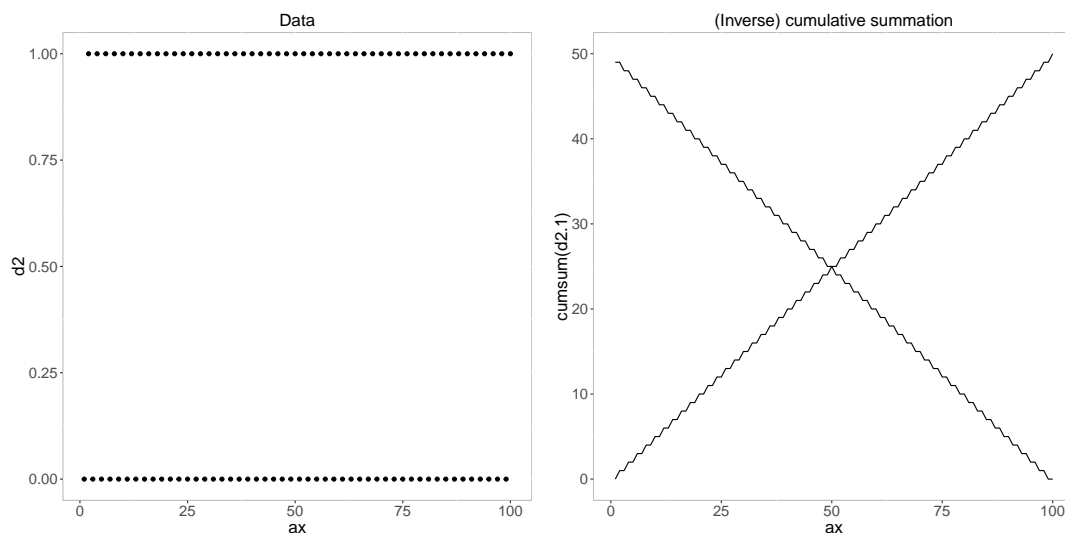
Predictions classified using a threshold of 0.5.

Figure 1.8: Switch-point identification using the intersection of cumulative summation of 'Do not Participate' and inverse cumulative summation of 'Participate' outcomes for each approach.



Cumulative summation of 1's and inverse cumulative summation of 0's using example data.

Figure 1.9: Example of a perfect switch-point.

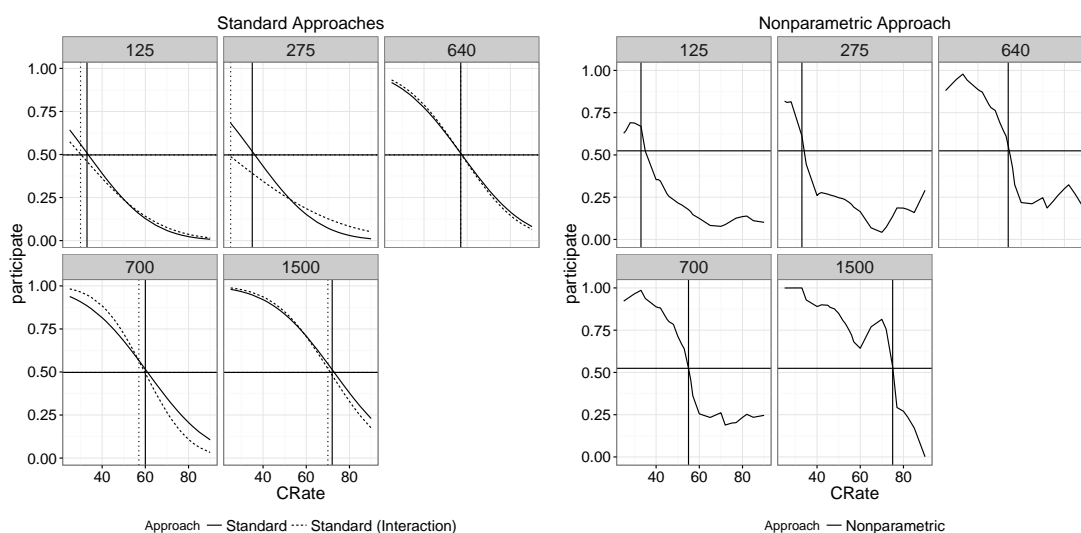


Cumulative summation of 1's and inverse cumulative summation of 0's using example data.

Figure 1.10: Example of absence of a switch-point.

1.5.2 Identifying candidates using predicted values: Youden's optimal J

For each of the approaches to smoothing taken in Section 1.4.2, candidate switch-points can be identified by mapping the single Youden's Optimal J (YOJ) of the approach onto the *CRate* for each level of *income*. Thus five candidate switch points are identified for each smoothing approach. The results of the Empirical approach are omitted here since these cross the YOJ value in multiple locations, leading to multiple values of *CRates* as candidate switch-points. Multiple intersections do not occur with the parametric approaches by design, and rarely occur under the Nonparametric approach. In cases where this does occur the median of the candidates identified is used. Figure 1.11



Horizontal lines indicate optimal Youden's J value, vertical lines indicate intersection with the profile of predictions.

Figure 1.11: Identification of switch-points by mapping optimal Youden's J values to contribution rates.

illustrates the method for the Standard approach both with and without the interaction term in the first pane and for the Nonparametric approach in the

second pane. The horizontal lines represent the optimal Youden's J values, and the vertical lines indicate the *CRate* where the YOJ intersects the predictions of a particular approach. The intersection point represents a candidate switch-point since all predictions to the left can be sorted as 'Participate' and all to the right as 'Do not Participate'. Following this sorting, the resultant candidate switch-points can be compared to those predicted by the theoretical model. Table 1.13 summarizes the candidate switch-points identified using this method. If the predictions do not intersect the YOJ value (as in the 275 *income* level for the Standard approach with interaction) the candidate switch-point is recorded as the lowest *CRate* which occurred in the observations.

Income	Standard	Standard (Interaction)	Nonparametric
125	33	30	33
275	35	25	33
640	57	57	54
700	60	57	55
1500	72	70	75

Table 1.13: Switch-points identified using the optimal Youden's J value.

The results for OYJ are very similar across smoothing strategies. The two Standard approaches return identical results for the 125, 700 and 1500 income levels, and very similar results for the remaining 275 and 640 income levels. The results of the Nonparametric approach identify candidate switch-points which are much more clearly reflected in the plotted predictions than in the Standard approaches. Since this method provides a result regardless of the steepness of the change in the predictions, no information is provided about the merit of the candidate switch-points identified.

1.5.3 Identifying candidates using predicted values: maximum absolute gradients

Because switch-points are substantively large changes in predicted participation decisions ($\widehat{participate}$) over a very small range of $CRate$ at each level of $income$ we can use the gradients of each of these approaches to compare the relative merits of each of the candidate switch-points.

A gradient is simply the rate of change in predicted probability at a particular $CRate$ and $income$. Gradients can provide information about the relative intensity of the candidate switch-points both across and within approaches. Comparing across approaches, the sizes of the gradients can offer support, or lack thereof, with larger gradients indicative of a stronger candidate. Within an approach the gradients for each $income$ level are a measure of the relative strength of the candidate switch-points, which will be discussed in this section. The details of the calculations are provided in Appendix Section 1.J. The results for the Standard approaches are the smooth lines plotted in Figure 1.12 while the Nonparametric gradients are the jagged lines. For the Standard approach the gradients become less negative as $CRate$ increases because the predictions decline at a decreasing rate in a smooth manner. The gradients of the Nonparametric approach reflect the less smooth nature of the predictions with sharp downward points representing steep changes in the predictions. As previously mentioned, steep changes in the predictions are indicative of switch-points. Using the Nonparametric approach rather than the Standard approach, switch points are unmasked and the gradients point directly to the candidates. Table 1.14 reports the $CRates$ indicated by the largest gradients in absolute value. For the Standard approaches these are simply the first $CRates$ encountered due to the nature of the predictions, while for the Nonparametric approach clear switch-points within the predictions are indicated.

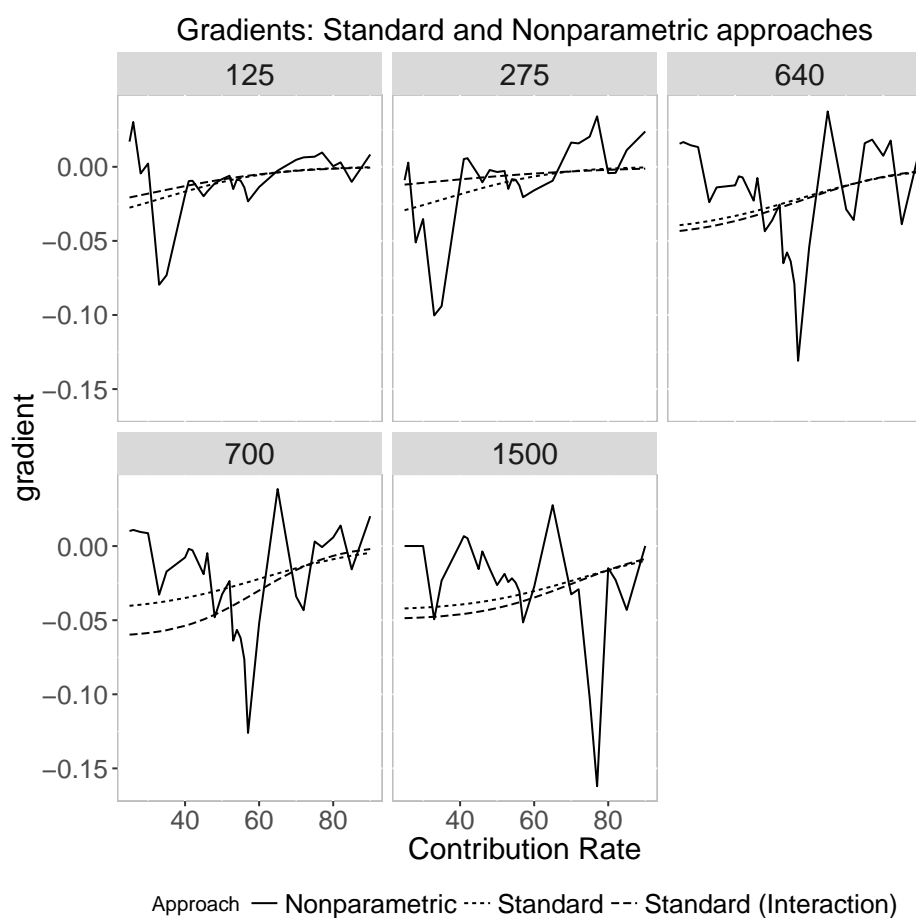


Figure 1.12: Gradients by contribution rate and income for each approach.

Income	Nonparametric	Standard	Standard (Interaction)
125	33	25	25
275	33	25	25
640	57	25	25
700	57	25	25
1500	77	25	25

Table 1.14: Switch-points identified using the maximum absolute gradient.

1.5.4 Comparison of candidate switch-points

Thus far five methods for identifying candidate switch-points have been proposed: Three for the observations alone and two for the predicted probabilities *participate*. The cumulative and inverse cumulative sum intersection (CSM) method is applied to the observations directly (See Appendix Section 1.I for the details of these methods). The Youden's optimal J (YOJ) mapping, and the maximum absolute gradient (MAG) were applied to the predictions of each approach. While the CSM method can be used on the observations without transformation, as well as any of the predictive approaches, it does not differentiate candidates from observations with an obvious discrete change from non-candidates and so is less attractive than the other options. The YOJ method can be applied only to predictions which are smoothed. This disqualifies the application of the YOJ method to the Empirical approach since the YOJ value intersects the predicted probabilities multiple times leading to multiple *CRate* candidates. For the Standard and Nonparametric approaches the YOJ identifies at least one threshold for optimally sorting predictions into 'Participate' and 'Do not Participate' categories regardless of the degree of smoothing, and so cannot discriminate approaches which exhibit steep changes in the predictions from gentle slopes.¹³ The MAG method can be applied also only to adequately smoothed predictions¹⁴ and identifies the point at which the largest change in the predictions occurs. This method is the more direct for locating candidates and evaluating their relative merits because both the location of the change in terms of *CRate* and the value of the

¹³If multiple values are encountered during the bootstrap process the median of the candidate *CRates* is used. This is an issue only for the Nonparametric approach since the Standard approaches are uniformly downward sloping.

¹⁴Inadequately smoothed predictions may result in non-unique maximum absolute gradients.

the gradient at this point (a measure of the degree of the change in predictions) are provided.¹⁵

1.6 Results

1.6.1 Bootstrapping procedure

The main purpose of the identification and assessment of candidate switch-points is to compare the observations gathered in the experiment with the predictions of the theoretical model of Section 1.3. Confidence intervals are constructed using the following simple nonparametric bootstrapping approach:

1. An identifier is applied to demarcate independent observations for each group of 5 participants in each period of the experiment. This leads to 100 unique identifier values.
2. A sample, with replacement, of 100 observations of the identifiers is taken.
3. For each identifier, the set of 5 triads of observations of *Income*, *CRate*, *participate* are pulled into the bootstrap sample (i.e. a new set of 500 observations). From this re-sample of 500:
 - Smooth the new set of observations using the approach of choice (Standard or Nonparametric) and record the results.
 - Identify candidates using the CSM, YOJ or MAG methods and record.
4. Repeat steps 2 and 3 1000 times.

¹⁵A method for constructing a measure of strength of the YOJ candidates using both YOJ and MAG information is provided in Appendix Section 1.K.

5. Record the 5th and 95th percentiles of the observations and candidates as the upper and lower bounds of confidence intervals.

This is the same approach taken for constructing all confidence intervals unless otherwise reported.

1.6.2 Observations vs theoretical predictions

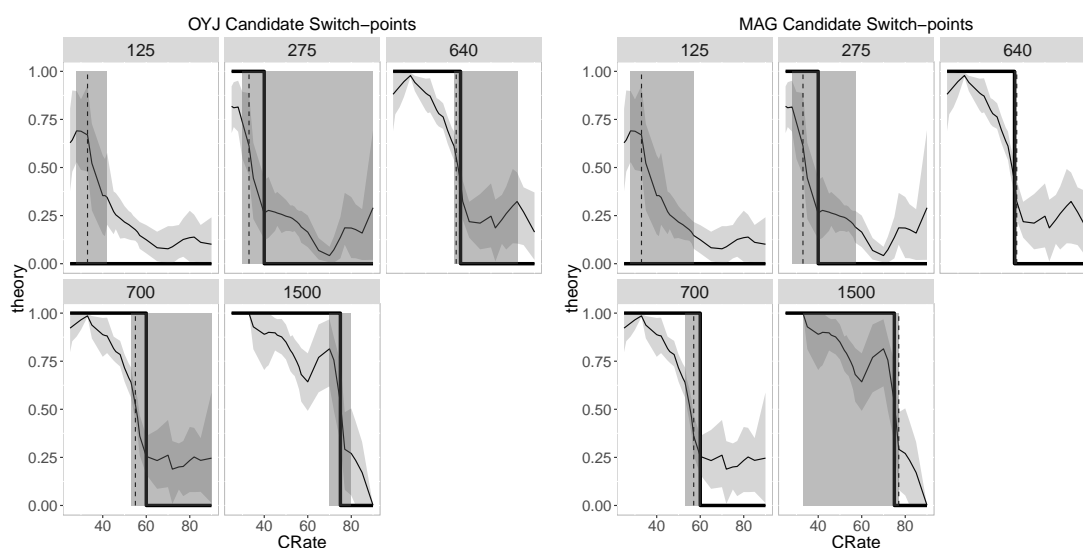
Approach	Income	Theoretical	MAG	Lower	Upper	OYJ	Lower	Upper
Standard	125	20	25	25	26	33	25	35
Standard	275	35	25	25	26	35	25	25
Standard	640	56	25	25	26	57	50	57
Standard	700	58	25	25	26	60	53	65
Standard	1500	75	25	25	26	72	60	75
Standard (Interaction)	125	20	25	25	26	30	25	40
Standard (Interaction)	275	35	25	25	26	25	25	35
Standard (Interaction)	640	56	25	25	26	57	50	60
Standard (Interaction)	700	58	25	25	26	57	53	65
Standard (Interaction)	1500	75	25	25	26	70	60	75
Nonparametric	125	20	33	28	57	33	28	42
Nonparametric	275	35	33	28	57	33	30	90
Nonparametric	640	56	57	56	57	54	53	82
Nonparametric	700	58	57	53	60	55	53	90
Nonparametric	1500	75	77	33	77	75	70	80

Maximum absolute gradient (MAG) and optimal Youden's J (OYJ) switch-point identification. Upper and lower bounds of the bootstrapped 90 percent confidence interval.

Table 1.15: MAG and OYJ method switch-point candidates and confidence intervals for each approach.

The confidence intervals for the candidate switch-points identified using the MAG and YOJ methods are presented in Table 1.15. For the Nonparametric approach, all candidates identified using the MAG method, except the 125 *income* level, have associated confidence intervals which include the theoretical prediction. This indicates that the observed switch-point is in agreement with the prediction of the theoretical model. Figure 1.13 illustrates. In a number of cases agreement with the theoretical predictions is driven by the finding of wider confidence intervals. The 640 *income* level strongly supports the theoretical prediction. Similar results hold for the candidates identified using the YOJ method, though the confidence intervals are generally wider, indi-

cating a lower degree of precision associated with this method of identifying switch-point candidates. This combination of results suggests that the theoretical model cannot be rejected as an explanation of behaviour for all but the lowest income group. For the Standard approaches illustrated in Figures 1.14



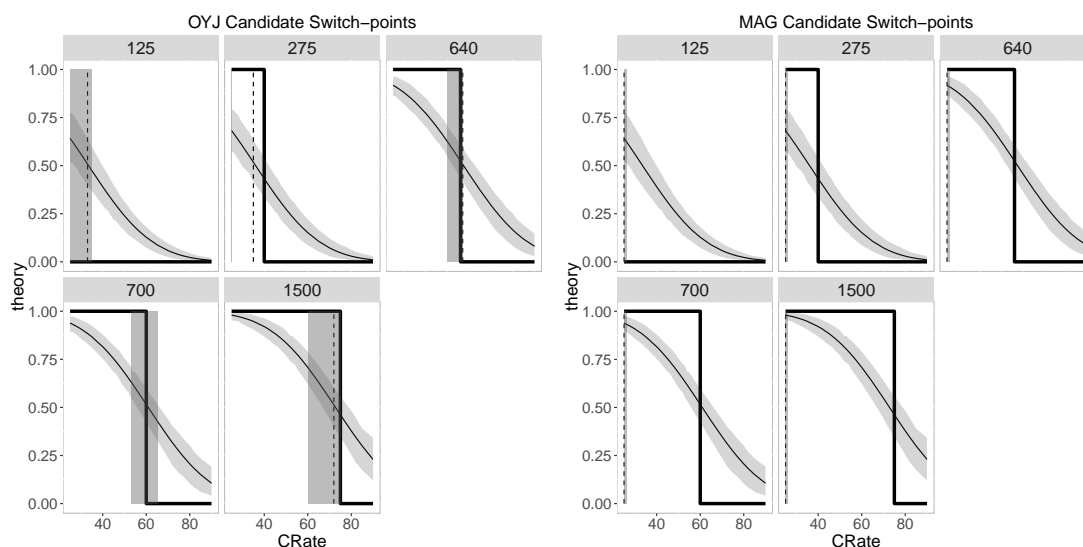
Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 1.13: Candidate switch-points of the MAG and OYJ methods for the Nonparametric approach.

and 1.15, the results of the MAG and YOJ method differ substantially. The MAG confidence intervals suggest a complete rejection of the coincidence of the candidates with the predictions of the theoretical model. All candidates are identified as the first *CRate* encountered and the associated confidence intervals are very small. While the small confidence intervals might be suggestive of a high degree of precision, it is precision associated with an inconsistent candidate, a result of the particular specification of the Standard technique. This is further evidence against the use of the Standard method of smoothing. The YOJ candidates are closer to the theoretical predictions for all except the lowest *income* level, suggesting that the theoretical model partially

explains the results. The confidence intervals are wider than those using the MAG method and contain the estimates of the theoretical model in all but the two lowest *income* instances.

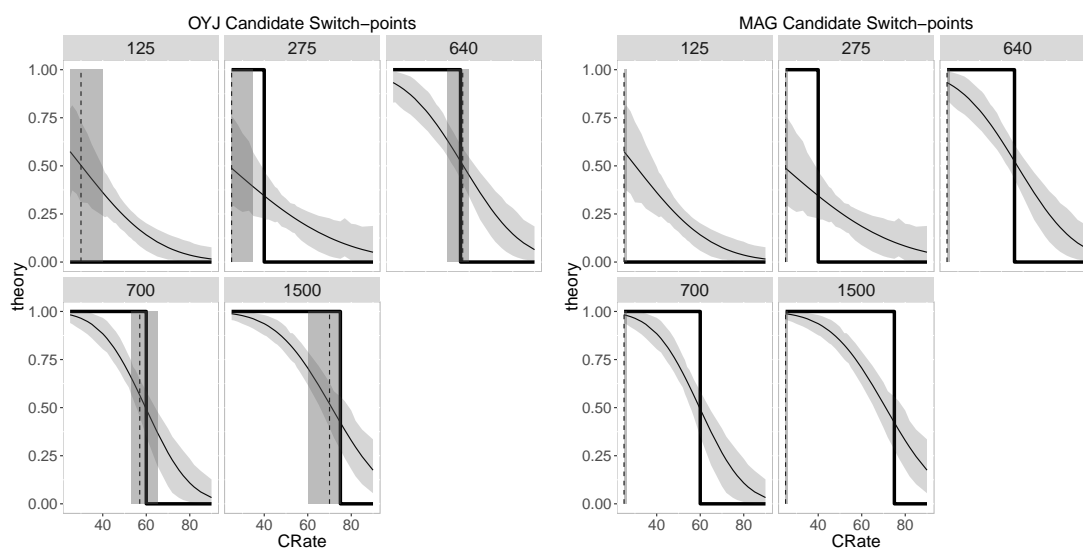
For the Standard approach with interactions the results are similar to the Standard without interactions and raise an important consideration for the treatment of predictions which do not intersect the optimal Youden's J value. In this case the 275 level of *income* predictions do not intersect the Youden's optimal J value and therefore have no associated *CRate*. Rather than leaving this case undefined the *CRate* can be substituted as the minimum *CRate* which occurs in the data set, or as a 0. Here the minimum value in the data set is used.



Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 1.14: Candidate switch-points of the MAG and OYJ methods for the Standard approach.

Overall, for the Standard approaches the more accurate MAG method for identifying candidate switch-points rejects the notion that the identified candidates are consistent with the theoretical model, while the less precise YOJ



Theoretical cutoffs bold lines, predictions solid lines, and candidate switch-points dotted lines. Bootstrapped 90 percent confidence intervals shaded in grey. Maximum absolute gradient (MAG) and optimal Youden's J (YOJ) switch-point identification methods.

Figure 1.15: Candidate switch-points of the MAG and OYJ methods for the Standard approach with interactions.

method identifies candidates which are closer to the theoretical predictions. Similar to the CSM approach, under the YOJ approach a switch-point can be suggested even in the case of predictions which exhibit a constant decline rather than a rapid change resembling a switch point.

1.7 Conclusion and discussion

In this paper observations of an economic experiment were compared to predictions from the theoretical model on which the experiment was based. The theoretical model predicts a discrete change in the binary outcome, forming a 'step' pattern. Three approaches to smoothing observations were investigated, the Nonparametric approach was shown to be the preferred approach by a number of criteria and, within the predicted *participate*, the MAG method the most effective for identifying candidate switch-points. While the

candidate switch-points of each method are similar across approaches, a visual inspection of the gradients demonstrates that the Nonparametric candidates are better at capturing substantive changes in predictions than the Standard candidates.¹⁶ Finally, Nonparametric bootstrapped confidence intervals were constructed around the candidate switch-points and the match between the identified switch points and cutoffs suggested by the theoretical model evaluated.

While the Standard approach with interactions suggested rejection of the coincidence of the candidate switch-points with the theoretical cutoffs for the two lowest levels of *income* these results are likely invalid. Based on comparison to the confidence intervals of the Nonparametric approach, the Standard approach appears to offer a high level of precision around incorrect estimates. Combined with results which suggest that the parametric models are misspecified and weakly indicative of switch-points, inference should not be made about the behaviours captured in the experimental data based on the Standard approaches.

The Nonparametric approach and MAG method extend naturally to other fields in which analysts wish to confront observed data with an externally determined cutoff for the purpose of evaluating the effectiveness of such a cutoff. The potential policy applications of this framework are diverse. For example, medical practice guideline evaluation could be improved by the application of the more robust Nonparametric approach combined with the MAG method for identifying switch points in observed practitioner behavior. Firm entry and exit decisions over a range of prices can be analysed using this method, as can the uptake of public programming over a range of incomes of individuals. The flexibility of the smoothing also allows for eval-

¹⁶The statistic E was proposed in Appendix Section 1.K as a means of condensing the YOJ results into a measure of the intensity of each candidate switch-point for comparison across and within models and potentially across data sets.

uation of more complex guidelines. In addition, by smoothing the observations themselves, rather than the observations which are in agreement with the theoretical predictions, the smoothing approach, switch-point identification method and the guideline can be easily compared against alternative specifications visually. This intuitive visual appeal of the results provides the added benefit of fostering the communication of the results of evaluations undertaken using this method with diverse audiences.

References

- Aitchison, J. and Aitken, C. G. G. (1976). "Multivariate binary discrimination by the kernel method". In: *Biometrika* 63.3, pp. 413–420.
- Akaike, H. (1974). "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on* 19.6, pp. 716–723.
- Buckley, N., Cuff, K., Hurley, J., Mestelman, S., Thomas, S., and Cameron, D. (2015). "Support for public provision of a private good with top-up and opt-out: A controlled laboratory experiment". In: *Journal of Economic Behavior & Organization* 111.0, pp. 177–196.
- Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.
- Cragg, J. G. and Uhler, R. S. (1970). "The demand for automobiles". In: *The Canadian Journal of Economics/Revue canadienne d'Economie* 3.3, pp. 386–406.
- Epanechnikov, V. A. (1969). "Non-Parametric Estimation of a Multivariate Probability Density". In: *Theory of Probability & Its Applications* 14.1, pp. 153–158. DOI: 10.1137/1114019. eprint: <http://dx.doi.org/10.1137/1114019>. URL: <http://dx.doi.org/10.1137/1114019>.
- Fawcett, T. (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.
- Fernihough, A. (2014). *mfX: Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs*. R package version 1.1. URL: <http://CRAN.R-project.org/package=mfX>.
- Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

- Hall, P., Racine, J. S., and Li, Q. (2004). "Cross-validation and the estimation of conditional probability densities". In: *Journal of the American Statistical Association* 99, pp. 1015–1026.
- Hayfield, T. and Racine, J. S. (2008). "Nonparametric Econometrics: The np Package". In: *Journal of Statistical Software* 27.5.
- Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.
- McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*. Institute of Urban and Regional Development, University of California.
- R Core Team (2015a). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/>.
- Racine, J. S. and Li, Q. (2004). "Nonparametric estimation of regression functions with both categorical and continuous data". In: *Journal of Econometrics* 119.1, pp. 99–130.
- Ramalho, E. A. and Ramalho, J. J. S. (2012). "Alternative Versions of the RESET Test for Binary Response Index Models: A Comparative Study*". In: *Oxford Bulletin of Economics and Statistics* 74.1, pp. 107–130. ISSN: 1468-0084.
- Ramsey, J. B. (1969). "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 350–371.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.
- Stone, C. J. (1977). "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4, pp. 595–620.
- Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. New York: Psychology Press.

Youden, W. J. (1950). "Index for rating diagnostic tests". In: *Cancer* 3.1, pp. 32–35. ISSN: 1097-0142.

1.A Frequency of observations

	25	26	28	30	33	35	40	41	42	45	46	48	50	52	53	54	55	56	57	60	65	70	72	75	77	80	82	85	90
125	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
275	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
640	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
700	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2
1500	2	1	1	3	1	2	4	2	1	2	1	1	23	1	4	1	9	5	3	9	3	7	1	2	1	6	1	1	2

Table 1.A.1: Number of observations by income and contribution rate

1.B Standard approach

The index function (the $X'\beta$) is described by:

$$X'\beta = \alpha + \beta_1 CRate + \beta_2 \mathbb{I}_{275} + \beta_3 \mathbb{I}_{640} + \beta_4 \mathbb{I}_{700} + \beta_5 \mathbb{I}_{1500}, \quad (1.B.1)$$

where $CRate$ and $income$ are explanatory variables, $\beta_2, \beta_3, \beta_4$, and β_5 are associated with indicators, $\mathbb{I} = 1$, for the sub-scripted income level and $\mathbb{I} = 0$ otherwise.

In order to interpret the coefficients in Table 1.3 as probabilities the index function must be distributed according to a cumulative normal distribution for the probit model. For example, the probability of a subject with an income of 125 choosing to participate can be described by:

$$\begin{aligned} \Pr(\text{participate} = 1 | \text{income} = 125, CRate) &= \Phi(\alpha + \beta_1 * CRate) & (1.B.2) \\ &= \Phi(1.438 - 0.043 * CRate). \end{aligned}$$

The model is estimated by maximum likelihood using iteratively weighted least squares.

1.C Standard approach with interaction

The probit model is updated to include interaction terms:

$$\begin{aligned} X'\beta = & \alpha + \beta_1 CRate + \beta_2 \mathbb{I}_{275} + \beta_3 \mathbb{I}_{640} + \beta_4 \mathbb{I}_{700} + \beta_5 \mathbb{I}_{1500} + \\ & \beta_6 \mathbb{I}_{275} * CRate + \beta_7 \mathbb{I}_{640} * CRate + \beta_8 \mathbb{I}_{700} * CRate + \beta_9 \mathbb{I}_{1500} * CRate, \end{aligned} \quad (1.C.1)$$

which is again, distributed according to $\Phi(\cdot)$ in order to interpret the results as probabilities. The \mathbb{I} represent indicator variables equal to 1 at the sub-scripted levels of *income* and 0 otherwise, as before.

So now, for example, the probability of a subject with an *income* level of 275 choosing to participate can be described by:

$$\begin{aligned} \Pr(Y = 1 | income = 275, CRate) &= \Phi(\alpha + \beta_1 * CRate + \beta_2 \mathbb{I}_{275} + \beta_6 \mathbb{I}_{275} * CRate) \\ &= \Phi(1.088 - 0.036 * CRate - 0.502 + \\ & \quad 0.011 * CRate). \end{aligned} \quad (1.C.2)$$

1.D Coefficients of determination

Often a coefficient of determination is used as a measure of goodness of fit, with various pseudo- R^2 calculations for predictions not calculated by ordinary least squares, as is the case here.

For example, the pseudo- R^2 described by McFadden (1973) is:

$$R^2 = 1 - \frac{\ln \hat{L}(M_{FULL}) - K}{\ln \hat{L}(M_{INTERCEPT})}, \quad (1.D.1)$$

where \hat{L} is the estimated likelihood, K the number of parameters, M_{FULL} is the full model and $M_{INTERCEPT}$ is the model with only an intercept included. The value of this statistic is 0.24, which is indicative of a good model

fit according to Louviere (2000).¹⁷ By this metric alone it is plausible that this model sufficiently describes the data. Another option is the R^2 of Cragg and Uhler (1970), which is an adjusted version of the R^2 value of Cox and Snell (1989), and is described by :

$$R^2 = \frac{1 - \left\{ \frac{L(M_{INTERCEPT})}{L(M_{FULL})} \right\}^{\frac{2}{N}}}{1 - L(M_{INTERCEPT})^{\frac{2}{N}}}, \quad (1.D.2)$$

where L is the log likelihood, M_{FULL} and $M_{INTERCEPT}$ are the same as previously defined and N is the number of observations in the data set. This statistic takes on values between 0 and 1, with 0 indicating a poor fit and 1 a very good fit. The value here is 0.4. Another R^2 which is independent of the particular approach used is the adjusted count R^2 described by:

$$R^2 = \frac{Correct - n}{Total - n}, \quad (1.D.3)$$

where $Correct$ are the number of predicted outcomes ≥ 0.5 , $Total$ is the total number of observations, and n is the count of the most frequent outcome. Here it has a value of 0.5. Each of these measures has a slightly different interpretation so they are not directly comparable, however it is not uncommon to see values in these ranges cited as indicative of good model fits.

1.E Bandwidth selection

The benefits of least squares cross validation are described in detail by Hall et al. (2004). This method minimizes the weighted integrated squared error

¹⁷ Louviere (2000) suggests that a value between 0.2-0.4 indicates a very good model fit.

described by:

$$ISE = \int \hat{g}(y|x) - g(y|x)^2 \mu(x) M(x^c) dx dy, \quad (1.E.1)$$

where $M(x^c)$ is a weight function which can then be minimized by least squares cross validation, described by:¹⁸

$$CV_{g_0} = \int \frac{[(\hat{f})(x, y) - \hat{g}(x)g(y|x)]^2 M(x^c)}{\mu(x)^2} dx dy$$

and

$$CV_{g_0}(h_0, h, \lambda) = n^{\frac{-q}{(q+4)}} \chi_g(a_0, a, b) \quad (1.E.2)$$

This approach has the distinctive benefit that irrelevant elements of X can be smoothed completely out of the regression.

An alternative to least squares cross validation is to use maximum likelihood cross validation, which can tend to oversmooth for fat-tailed distributions:

$$\hat{g}_{-i}(Y_i|X_i) = \frac{\hat{f}_{-i}(X_i, Y_i)}{\hat{m}_{-i}(X_i)}. \quad (1.E.3)$$

1.F Is *income* relevant?

Variable	Least Squares	Maximum Likelihood	Least Squares Without Income	Upper Bound
Participate	0.0000	0.0512	3e-04	0.5
Income	0.9926	0.9923	-	1
CRate	3.1305	3.1305	2.8407	inf

Table 1.F.1: Bandwidths generated using least squares cross validation (with and without income) and maximum likelihood cross validation.

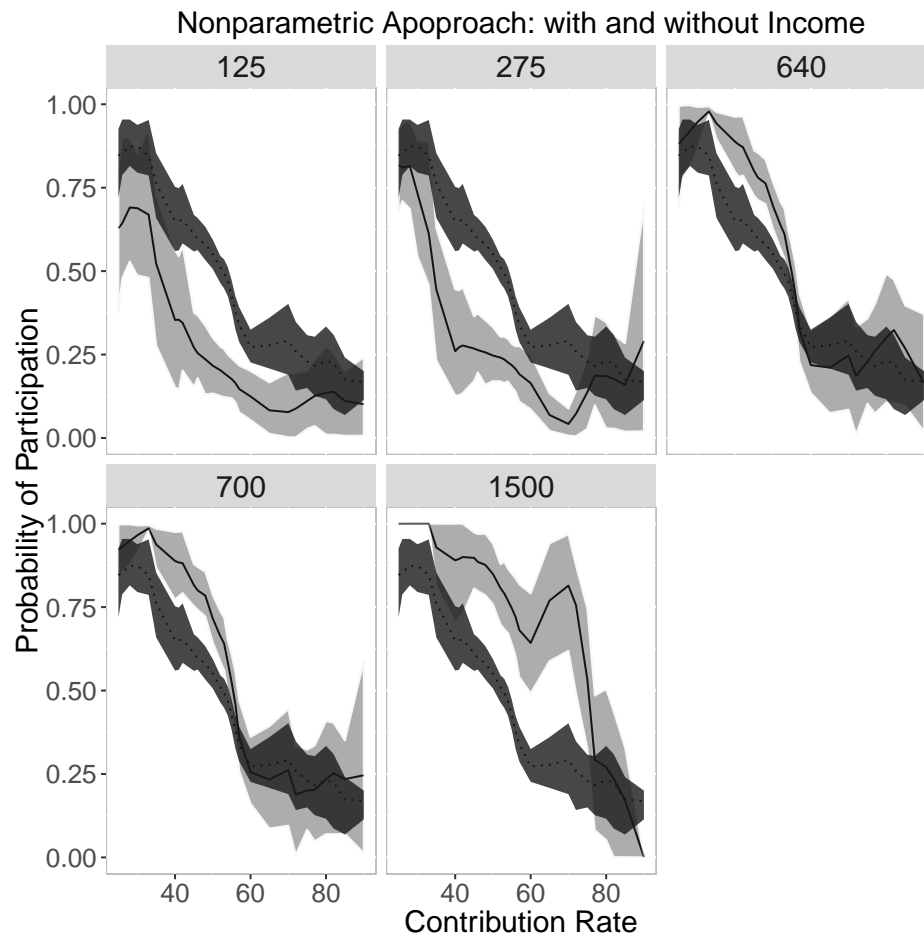
It is worth noting that even when using the maximum likelihood strategy *income* is not completely smoothed out, as would be indicated if the bandwidth

¹⁸See Li and Racine (2007) chapter 5, pp 157-160 for more details about this function

were equal to 1.00. As noted in Racine and Li (2004) there may be times in which a smoothing parameter may be close to its upper bound but not be irrelevant. As a check for the relevance of *income* the least squares cross validation bandwidth selection routine is run for the conditional probability of *participate* on *CRate* alone. The column 'Least Squares Without Income' in Table 1.F.1 shows the resulting smoothing parameters when *income* is omitted from the estimation. The bandwidth for *participate* is nearly identical to the least squares bandwidth when *income* is included, while the bandwidth of *CRate* is lower, indicating that more of the variation in *CRate* is now explaining variation in the conditional probability of participation.

Comparing the Adjusted Count R^2_{AC} of each estimation described by Equation 1.D.3 assists in clarifying the relevance of *income* in estimating the conditional probability of *participate* because the value is 0.559 in the case where *income* is included and 0.284 when it is not. These results support the continued inclusion of *income* despite having a smoothing parameter near its upper bound. If *income* were truly irrelevant these R^2 values should be nearly identical. In what follows, the bandwidths resulting from the least squares cross validation procedure with probability of participation conditional on both *income* and *CRate* will be used.

Figure 1.F.1 plots the results for the Nonparametric approach both with and without *income* as an explanatory variable, where without *income* the predictions remain constant across the five panes of the figure. It is clear that the predictions of the model without *income* differ substantially from the model with *income*, supporting the inclusion of *income* since if *income* were truly irrelevant the results should be identical.



Solid lines and lighter gray confidence bands include income as an explanatory variable. Dotted lines and darker gray confidence bands exclude income as an explanatory variable. Confidence intervals are bootstrapped 90 percent bounds.

Figure 1.F.1: Predicted probability of participation by contribution rate (with and without income) using the Nonparametric approach.

1.G Receiver operator characteristics curve

The True Positive Rate (TPR) is the percentage of predicted positive decisions which match the observed positive decisions. This is expressed as:

$$TPR = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}, \quad (1.G.1)$$

and is a measure of the ability of the theoretical model to correctly predict a positive response. Taking an example from the medical literature where such

methods are frequently used, this measures the ability of a diagnostic test to correctly identify a disease in a person who indeed has the disease.

Increasing the ability of a model to correctly predict positive outcomes is typically done at the expense of increasing the False Positive rate: of incorrectly predicting illness in a healthy patient. This measure is the False Positive Rate (FPR) and is expressed:

$$FPR = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}} \quad (1.G.2)$$

Combining the TPR and FPR, ROC curves capture the responsiveness of a particular estimation strategy by varying the threshold from 0 to 1, recalculating the confusion matrix and plotting the TPR against the FPR. Figure 1.7 in Section 1.4.4.2 above plots the results for the models discussed in Section 1.4.2. Since the measures constructed from a confusion matrix do not depend on underlying assumptions about the model from which the predictions were generated, the results are comparable regardless of the source of the predictions. When the threshold is 0 all predictions are classified as 0's and thus no positives (1's) are identified at all. When the FPR is 0 it must also be the case that the TPR is zero since there are no positives identified at all. This point occurs at the origin of Figure 1.7. When the threshold is 1 all predictions are classified as 1's, no negatives are identified, and the TPR and FPR are thus both 100%. This point occurs in the North East corner of Figure 1.7. More concave ROC curves are indicative of predictions which are closer to the observed data. A curve going from (0,0) to (0,100), to (100,100) represents a situation in which the predictions perfectly match the outcomes regardless of the threshold used. The 45° line on the other hand is representative of a situation in which varying the threshold causes an exactly proportional increase in the FPR and TPR indicating that the results of the predictive strat-

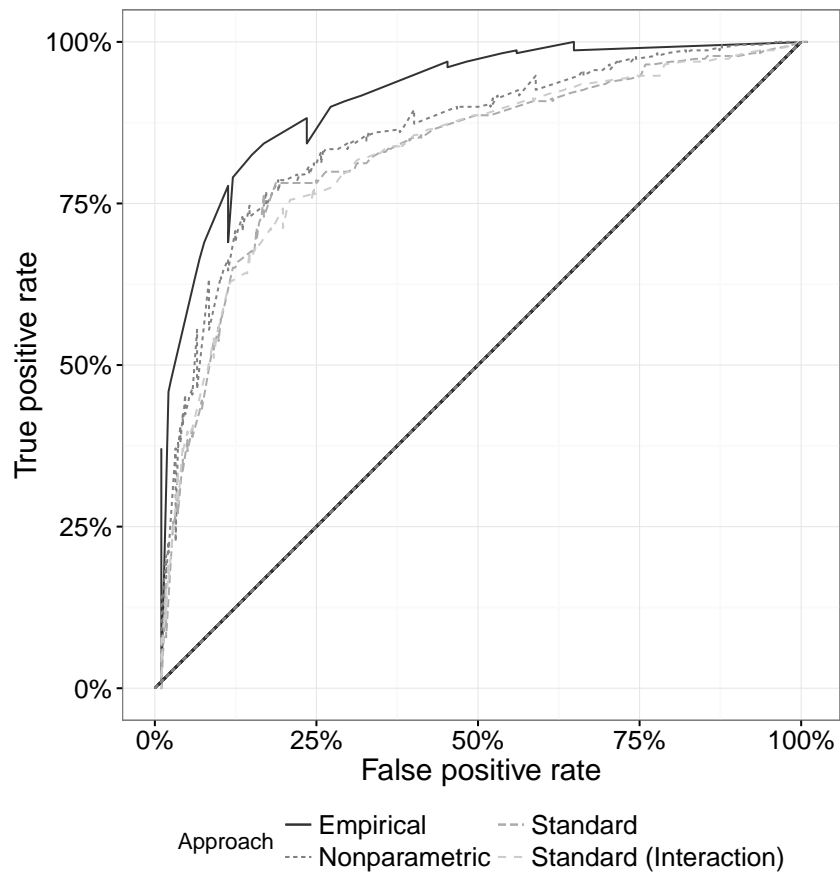


Figure 1.7: Receiver operator characteristics curves for each approach. (repeated from page 36)

egy do not describe the data at hand.

1.H Cumulative summation intersection

Cumulative summation is the sum of the frequencies of an outcome up to a particular *CRate* and *income*. Here it is defined as the sum of the 0 observations up to a particular $CRate(j)$ for each $income(k)$:

$$CS_{jk} = \sum_1^j \sum_i^N I\{(y_i | CRate = j, income = k) = 0\} \quad (1.H.1)$$

As more observations are encountered the cumulative summation will either increase or stay the same, depending on the outcome of interest. The inverse of the cumulative summation is the cumulative frequency of an outcome up to a particular *CRate* and *income* subtracted from the maximum value of the cumulative summation.

$$ICS_{jk} = \sum_1^J \sum_i^N I\{(y_i | CRate = j, income = k) = 1\} - \sum_1^j \sum_i^N I\{(y_i | CRate = j, income = k) = 1\} \quad (1.H.2)$$

1.I Identifying candidates using observations: search method

Another way the raw observations can be used to identify candidate switch-points is via searching for a candidate which maximizes a criteria. This method is convenient because it does not require smoothing, but it may remain sensitive to noise.

To carry out a search for an optimal candidate, for each level of income:

1. Choose a switch-point candidate,
2. Create a 'pseudo-step': define all *CRates* equal or less than the candidate

as 'Participate' and all *CRates* greater as 'Do not Participate'.

3. Construct a confusion matrix of the observations and pseudo-step.
4. Calculate the J value.
5. Repeat the procedure for each possible candidate (every possible *CRate*).

The *CRate* associated with the maximum J value is then the 'optimal J' (*oj*) candidate. The results are presented in Table 1.I.1. The largest value of J is obtained at the 640 and 700 levels of *income* for which the thresholds are 53.

Income	Candidate	J
125	46	0.35
275	33	0.29
640	53	0.60
700	53	0.60
1500	55	0.43

Table 1.I.1: Candidate switch-points by search approach and optimal Youden's J method.

Similarly, maximizing the correct proportion instead of the value of J is another interpolation option. This method leads to multiple maxima in some instances. Table 1.I.2. In cases with multiple maxima the median value is reported as the candidate switch-point. The 640 level of *income* exhibits the highest proportion of correctly classified observations at the optimal threshold of 53.

Income	Candidate	Min	Max	Correct Proportion
125	35	33	40	0.10
275	33	33	33	0.26
640	53	53	53	0.57
700	55	53	57	0.50
1500	70	70	70	0.31

Table 1.I.2: Candidate switch-points by search approach and maximum correct proportion method.

Finally, the same procedure is replicated in Table 1.I.3, maximizing the value of Cohen's κ . Here the 640 level of *income* candidate is the strongest, exhibiting 60% agreement beyond chance at the optimal threshold of 53.

Income	Candidate	Cohen's kappa
125	46	0.36
275	33	0.38
640	53	0.60
700	53	0.58
1500	70	0.45

Table 1.I.3: Candidate switch-points by search approach and maximum Cohen's kappa method

Overall there is agreement between methods for locating an optimal threshold on income levels of 275 and 640. The search methods are convenient because they require no smoothing but do not lend well to intuitive graphical presentation of the results.

1.J Gradients

For the Standard model the gradient is calculated by taking the first derivative of the function $\Phi(X'\beta)$ and is defined as:

$$\frac{\partial Pr(Y = 1|X)}{\partial CRate} = \phi(X'\beta) * \frac{\partial(X'\beta)}{\partial CRate} \quad (1.J.1)$$

where $\Phi' = \phi$ is the derivative of the cumulative normal distribution. This leads to an equation into which the coefficients from 1.3 can be substituted and the exact predictions calculated for each value of *CRate* and *income* level encountered. For the 275 *income* level in the probit specification without inter-

action terms this amounts to:

$$\frac{\partial Pr(Y = 1|X)}{\partial CRate} = \phi(\alpha + \beta_1 * CRate + \beta_2 * D1_{income=275}) * \beta_2 \quad (1.J.2)$$

$$= \phi(1.438 + -0.043 * CRate + 0.109) * 0.109 \quad (1.J.3)$$

For the Nonparametric approach the gradient is simply computed as the derivative at every point in the predictions:

$$\frac{\partial \hat{g}(y|x)}{\partial x} \quad (1.J.4)$$

Approach	Income	MAG	Lower Bound	Upper Bound
Standard	125	-0.0218	-0.0430	-0.0186
Standard	275	-0.0237	-0.0429	-0.0203
Standard	640	-0.0218	-0.0539	-0.0301
Standard	700	-0.0243	-0.0542	-0.0310
Standard	1500	-0.0184	-0.0558	-0.0331
Standard (Interaction)	125	-0.0166	-0.0565	-0.0055
Standard (Interaction)	275	-0.0101	-0.0409	-0.0017
Standard (Interaction)	640	-0.0234	-0.0736	-0.0253
Standard (Interaction)	700	-0.0344	-0.1030	-0.0399
Standard (Interaction)	1500	-0.0189	-0.0721	-0.0316
Nonparametric	125	-0.0796	-0.1921	-0.0381
Nonparametric	275	-0.1003	-0.2180	-0.0410
Nonparametric	640	-0.1309	-0.1994	-0.0858
Nonparametric	700	-0.1261	-0.1924	-0.0834
Nonparametric	1500	-0.1620	-0.3307	-0.0675

Bounds are bootstrapped 90 percent confidence intervals.

Table 1.J.1: Gradients at the MAG candidate switch-points.

Approach	Income	OYJ	Lower Bound	Upper Bound
Standard	125	-0.0218	-0.0323	-0.0181
Standard	275	-0.0237	-0.0429	-0.0203
Standard	640	-0.0240	-0.0319	-0.0189
Standard	700	-0.0258	-0.0321	-0.0192
Standard	1500	-0.0198	-0.0331	-0.0190
Standard (Interaction)	125	-0.0166	-0.0404	-0.0055
Standard (Interaction)	275	-0.0101	-0.0327	-0.0017
Standard (Interaction)	640	-0.0260	-0.0422	-0.0155
Standard (Interaction)	700	-0.0372	-0.0568	-0.0224
Standard (Interaction)	1500	-0.0208	-0.0426	-0.0178
Nonparametric	125	-0.0796	-0.1568	0.0213
Nonparametric	275	-0.1003	-0.1788	0.0457
Nonparametric	640	-0.0578	-0.0977	0.0226
Nonparametric	700	-0.0620	-0.1103	0.0295
Nonparametric	1500	-0.1027	-0.1538	0.0165

Bounds are bootstrapped 90 percent confidence intervals.

Table 1.J.2: Gradients at the OYJ candidate switch-points.

1.K A strength measure for OYJ candidates

To construct a measure of the strength of a candidate switch-point which captures both the proportion of the largest change and the relative distance of the candidate to the maximum change in terms of $CRate$ the following are identified:

$$Gradient_M : \text{The true value of the largest gradient} \quad (1.K.1)$$

(when ranked by absolute value)

$$Gradient_Y : \text{The true value of the gradient at the candidate switch-point} \quad (1.K.2)$$

$$CRate_M : \text{The } CRate \text{ associated with the largest gradient} \quad (1.K.3)$$

$$CRate_Y : \text{The } CRate \text{ of the candidate switch-point} \quad (1.K.4)$$

The largest gradient is identified by ranking the gradients in terms of absolute value and selecting the maximum. The actual value of this gradient is recorded because in the Nonparametric approach a difference in sign from the candidate switch-point can be indicative of noise. This is in contrast to the

parametric model which imposes unidirectionality upon the predictions. If the candidate switch-point and the maximum gradient are the same or very close the data exhibit a strong switch-point since the optimal cutoff for sorting (the optimal Youden's J mapped into a *CRate*) accurately captures the location where the most substantive change occurs in the predictions. If the optimal Youden's J *CRate* and the *CRate* where the maximum gradient occurs are far apart then most of the changes in the predictions are occurring apart from the candidate, which should weaken the candidate's attractiveness. One way to concisely compare the maximum gradients and candidate switch-points is to take the percentage of the maximum gradient ($Gradient_M$) captured by the candidate ($Gradient_Y$) in terms of the relative distance of the associated *CRates*. If the $Gradient_M$ occurs close to the candidate then an area within the predictions exhibiting substantive change has been identified, if it occurs far from the candidate then either noise or an inappropriate model is indicated. The percentage deviation from the *CRate* at the maximum gradient from the optimal Youden's J *CRate* provides a unit-free measure of the spread between these two points. *CRates* located far apart may indicate noise or a model which fails to provide a switch-point. A simple measure of the relationship is suggested by the following:

$$GradientP = \frac{Gradient_Y}{Gradient_M} \quad (1.K.5)$$

$$CRateP = \left| \frac{CRate_Y - CRate_M}{CRate_Y} \right| \quad (1.K.6)$$

$$E = \frac{GradientP}{CRateP}, \quad (1.K.7)$$

where the absolute value of $CRateP$ is used because the direction of the the deviation is irrelevant. $GradientP \leq 1$ and may be negative if the maximum (in absolute terms) gradient differs in direction from the gradient identified at the candidate switch-point for a particular *income* level. The resulting mea-

sure E is akin to simply calculating the slope between the maximum gradient point and the one identified by the maximum Youden's J value, but avoids issues relating to the scale of $CRate$ and thus enables comparison across models and, potentially across data sets. Larger values of E indicate steeper slopes and thus stronger candidate switch-points. A perfect match of the candidate switch-point and maximum gradient provides the largest value E can take on ($\frac{100}{0} = \infty$).

Income	Standard	Standard (Interaction)	Nonparametric
125	4.12	6.00	Inf
275	3.50	Inf	Inf
640	1.96	1.98	7.94
700	1.82	1.93	13.53
1500	1.65	1.71	23.76

Table 1.K.1: Strength measure of candidate switch-points identified using the optimal Youden's J method.

Table 1.K.1 presents the results from each approach. The final values of E range from a low of 1.65 to a high of ∞ . In absolute terms, a value less than $|1|$ indicates particularly poor evidence of a switch-point since the relative distance between the YOJ and MAG candidate is greater than the percentage of the MAG gradient captured by the OYJ gradient. Values above $|1|$ are indicative of steeper relationships, as seen for the 125 and 275 *income* levels in Table 1.K.1, but these results offer little information given that the location of the maximum gradients are uniformly the first $CRate$ encountered, as discussed in the previous section.

The Nonparametric OYJ candidates are more reflective of steep changes than Standard OYJ candidates, as evidenced by the larger values of E . Nonparametric E values range from 7.94 to ∞ , with the minimum value exceeding the maximum value of the Standard approaches. For the 125 and 275 *income* levels the candidate identified by the Nonparametric YOJ is identical to the Nonparametric MAG candidate and the E of ∞ is ideal.

Chapter 2

A Standardized Method for the Evaluation of Adherence to Practice Guidelines

2.1 Introduction

Unnecessary use of medical care is a major concern for health care system administrators and patients. Not only can such care be costly, but it can also have negative health consequences. A recent article by Nelson (2015) highlights the cost and health consequences of unnecessary care in the setting of American hospitals, yet the issue extends to public health care systems and private practices alike. The Health Council of Canada (2009) cites that health care spending in Canada doubled between 1997 and 2007, with 48% of the increase attributable to increased use of health care services, but that the value of this greater service use is not fully understood. Practice guidelines are often advocated as a means of guiding practitioners in making appropriate care decisions to improve the quality of care, improve outcomes, and avoid unnecessary usage. Evaluation of adherence to such guidelines has, however, often

been done in limited study-specific settings involving large randomized clinical trials. This paper proposes a standardized methodology for evaluation of adherence to guidelines using existing administrative data records. The aim of this work is to facilitate greater undertaking of such evaluations, and therefore improve the information available upon which to base efficient care decisions. The method relies on a combination of basic analytical techniques and state-of-the-art data-driven analysis which leverages existing patient care data and has several attractive features. The techniques used require a minimum of subjective decisions on the part of the analyst and are widely applicable to contexts in which assessment of practice against a standard is required. The results are visually intuitive and can be easily communicated across diverse audiences. These features are particularly important for a standardized framework of guideline evaluation and should foster rapid adoption across health care sectors and hence improve the effectiveness of clinical practice guidelines in improving the efficiency and quality of care in health care systems.

Total expenditure on health care represents a substantial proportion of GDP among OECD member nations. Yet, in the US it is estimated that approximately 30% of care given to patients can be classed as 'unnecessary', meaning that it did not serve to improve patient health outcomes Nelson (2015). Understandably, reducing this share of unnecessary care and improving health system efficiency has become a priority concern for established research groups such as the Institute of Medicine (2006) and Health Council of Canada (2009), as well as smaller non-governmental groups such as the Lown Institute, exclusively concerned with 'Right Care' (The Lown Institute, 2016).

Medical practice guidelines are widely used to improve practitioner performance with the aim of improving patient care and promoting cost-effectiveness. Yet McGlynn et al. (2003) find that patients receive only about 50% of the

recommended care according to existing practice guidelines in the US. In Canada it has been found that nearly one quarter of seniors on public drug programs use a drug which has been identified as inappropriate (Canadian Institute for Health Information, 2014). The evaluation of practitioner adherence to evidence-based practice guidelines is particularly important in efforts to improve quality of care, and has been carried out in a variety of ways. While many studies rely on randomized controlled trials, these can be costly and time consuming to administer, as well as being highly specific to the particular clinical setting or health care guideline studied. Methods which make use of existing administrative patient care data have the potential to offer substantial insights about the performance of current health care systems and areas for improvement. Existing data can potentially be calibrated to offer insights on adherence in a much more readily accessible and constantly updated format, however. The standardized framework proposed in this paper provides a starting point for a richer analysis of adherence to clinical practice guidelines using existing data sources which can easily be adapted to a broad range of health system settings and guidelines.

Adherence to guidelines is addressed in various ways in the literature. Systematic reviews of the effectiveness of practice standards or factors influencing effectiveness have been conducted. Barbui et al. (2014) reviews practice standards in the setting of mental health care, Thomas et al. (1999) in the setting of professions allied to medicine. Flodgren et al. (2013) review standards to prevent device related infections, while Fiander et al. (2015) investigate standards to improve the use of electronic health records. Flodgren, Pomey, Taber, and Eccles (2011) collect studies about the effect of printed computerized reminders upon compliance with practice standards and Arditi et al. (2012) study the effect of inspections. O'Brien et al. (2007) survey the effect of educational outreach visits on compliance. Nearly all authors cite low quality

of evidence as key obstacles in drawing firm conclusions about the impact of interventions to improve practitioner compliance or health impacts of such interventions. An ongoing adherence monitoring system using the framework proposed in this paper would serve to harness existing data and overcome many issues related to lack of data encountered in previous studies, while offering a clear gauge of the level of adherence to a practice guideline. Outside of an RCT framework, adherence is also determined by modeling an outcome and comparing the result to the guideline. An example of this is presented by Askildsen, Holmås, and Kaarboe (2011), who model patient waiting times for surgery in Norway using a regression framework, comparing estimated wait times to ranges suggested by practice guidelines.¹ Dimakou, Parkin, Devlin, and Appleby (2009) examine the impact of government waiting time targets upon patient wait-times for surgery in the NHS. These authors take a hazard function approach to analysis and find that peaks in the probability of admission coincide with government mandated targets. The case by case variety in evaluation strategies encountered makes it difficult to compare and implement similar studies of adherence across jurisdictions, even for similar clinical processes, limiting the usefulness of such information in improving health system performance. The analytical strategy used in the framework proposed in this paper aims to provide a standardized, but flexible, approach to guideline adherence evaluation, therefore improving the ability of health system administrators to accurately assess and compare performance against established practice guidelines and to test and identify performance enhancing interventions.

The proposed framework for evaluation of adherence to guidelines serves

¹Adherence also enters analysis as an explanatory variable in a regression framework, for example by Andritsos and Tang (2014) who use a composite index of the fraction of agreement with clinical guidelines for a condition as a factor explaining the geometric mean of total in-patient stay for cardiac diagnoses. These authors argue that increased adherence leads to reduced resource use, but the effects are small.

to improve comparability across studies by suggesting a standard presentation of results. The first stage of analysis relies on a simple classification matrix approach which provides a basic overview of adherence and important additional information on non-adherence, splitting non-adherent cases into instances of under-treatment and over-treatment in accordance with a guideline. Summary measures of the classification matrix are provided and can easily be constructed from existing data given that results are reported in a classification matrix. In the second stage of analysis a state-of-the-art regression approach which relies on data and not on assumptions about functional form is used to generate a profile of estimated decisions. Adherence to the guideline is then assessed by comparing the estimated profile of decisions to the practice guideline by identifying key indicators within this profile. Because the estimation procedure is carried out independently of information about the guideline, the estimated profile of decisions serves as a test for the adherence to a guideline within a set of clinical observations. In addition, this lends to a naturally intuitive visual presentation of the results with the guideline and estimated profile of decisions plotted on the same axes.

Along with the analytical method proposed, this study contributes to the literature on adherence with an exploration of non-adherent decisions. Most studies of guideline compliance focus attention on the proportion of correct applications of a guideline. This work establishes a reference method and reasoning for including the evaluation of decisions which do not adhere to guidelines as well. Using a simple classification matrix, non-adherent decisions can be categorized as either over-treatment or under-treatment according to the guideline. Reducing over-treatment decisions represents potential future cost savings, while reducing under-treatment decisions represents potential future improvements in care. Non-adherent decisions are often not investigated in studies of guideline adherence. The reporting of this infor-

mation can provide important clues to understanding the effectiveness of a guideline.

In the language of Tugwell, Bennett, Sackett, and Haynes (1985), this paper represents a single element of a continuous system for assessing the value of any health care intervention on burden of illness. The framework proposed here relies on patient level administrative data to assess practitioner compliance with an established guideline. This work fits into the broader literature on health system assessment and should expand the range of health system monitoring to a variety of guidelines using existing administrative data as well as improving the comparability and reliability of the results.

The framework will be developed and illustrated with the assistance of an application to a data set consisting of the decisions of volunteer medical first responders (MFRs) to administer supplemental oxygen to patients encountered during regular service duties. Section 2.2 describes the data. Sections 2.3 and 2.5 present descriptions of each stage of the framework followed by applications to the data in Sections 2.4 and 2.6. The medical and economic impacts of the decisions made by MFRs are examined with the results indicating that overall, adherence is poor. There is a tendency towards over-treatment rather than failure to treat medically necessary cases. This effect is more pronounced for serious incidents. Fortunately, in economic terms the impact of these over-treatment decisions is low and in medical terms, potentially beneficial.

2.2 Data

An illustration of the proposed framework is provided using a data set consisting of the observed oxygen administration decisions of volunteer MFRs as they carried out regular duties in a large metropolitan area over the years

2010-2014. MFRs come from a variety of backgrounds, some with prior experience with oxygen administration and others without. Extensive mandatory training is provided at no cost to MFRs accepted into the organization. All MFRs are well informed that any deviations either below or above the standard of care practice guidelines outlined in their training qualify as a breach of duty and the legal ramifications of such breaches are discussed in depth.² Given the voluntary nature of the organization, additional group training sessions have been favoured to direct enforcement of guidelines through penalties imposed upon individual MFRs.

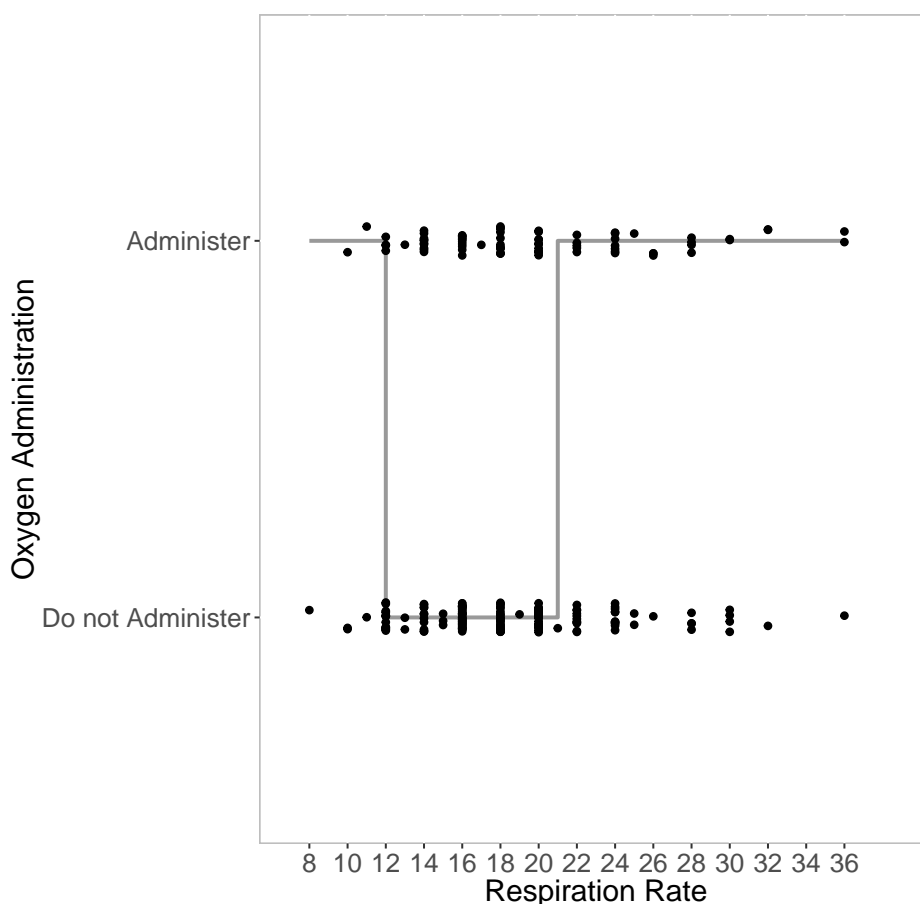
It is important to note that for legal purposes any care provided outside of the pre-defined scope of practice qualifies as a breach of duty. Thus, while in uniform with the organization a physician, nurse or paramedic is expected to perform to the standard of care defined by the organization and not that which their medical training might dictate. It is therefore expected that practice guidelines are strictly adhered to in the field. However, supplemental oxygen is rarely harmful for otherwise healthy patients and may even be used as a comfort measure to provide both patients and responders with the sense that they are doing something to improve a situation, regardless of medical benefit. For this reason non-adherence to the guideline is informative, and more specifically, the manner in which this non-adherence occurs. Administering oxygen which is not medically necessary is unlikely to expose a patient to a health risk, while not administering oxygen when the guideline dictates doing so may have substantial adverse medical consequences.³

The medical practice guideline in place was developed and issued as a stand-

²Responders are made aware that they may be called to testify in court for any treatment administered, that they can be represented by the lawyers of the organization, and that ultimately they carry the legal burden to correctly follow the standard of care outlined by the organization.

³It is worth noting here that patients refuse oxygen at times. Patient wishes are respected at all times.

ing order by the organization's Provincial Medical Director. Proper assessment of respiration rate and administration of supplemental oxygen is taught to all MFRs during training. The guideline requires MFRs to administer supplemental oxygen if an adult patient exhibits a respiration rate less than 12 or greater than 20 breaths per minute. Plotting the expected oxygen administration decisions over respiration rates ranging from 0-36 results in a straight line at 1 ('Administer') with a step down to 0 ('Do not Administer') at 12 and a step up to 1 ('Administer') at 20 and is illustrated in Figure 2.1. With perfect compliance the points, which represent the observed administration decisions (jittered slightly), would align with the horizontal portions of the line.



Solid lines represent the recommendation. Points represent observed decisions.

Figure 2.1: Recommended and observed oxygen administration decisions by respiration rate.

An anonymized set of 5 years worth of treatment decisions of volunteer MFRs form the data set used in this study. The international volunteer organization is sub-divided into national and provincial sub-units, and further sectioned into municipal level units. MFR training is standardized at the provincial level. The municipal level MFR unit covered by the data operates within a large Canadian metropolitan area, providing first-aid at local sporting events, concerts, festivals and public functions. Only fully certified MFRs are qualified to wear the full uniform of the organization and to provide first-aid. MFRs are also trained to complete a Patient Care Report (PCR) in the event of delivery of any form of first aid to a patient. PCRs are filed with the unit chief at the end of each duty shift.

The database includes a total of 898 medical-encounter records containing information on patient age, type of medical situation encountered by the MFRs, administration of oxygen, vital signs including respiration rate, and whether or not Emergency Medical Services (EMS) were contacted. All data were anonymized.⁴ Reports missing a year of birth of the patient were excluded (71 cases), as were the records of patients under the age of 18 (271 records) because a different guideline applies to these cases. As well, a single outlier which appeared to be a recording error in the respiration rate was dropped (1 record). Of these remaining 555 cases, 315 cases were missing a recorded respiration rate (these cases are discussed in more detail in Subsection 2.2.1. These exclusions resulted in 240 complete cases.

Table 2.1 displays the variables considered in the analysis of supplemental oxygen administration behaviour. The use of supplemental oxygen is recorded simply as 0 if it was not used and 1 if it was used. MFRs monitor vital signs throughout an incident, up to 5 times for a single patient. RR1 is respiration rate recorded at the first vital signs check. Respiration Rate values in the

⁴Reclassification ensured at least 5 observations per cell.

range of 12-20 respirations per minute are considered normal for adults. EMS refers to whether or not emergency services were called to the scene (0 if not called and 1 if called). This variable acts as an indicator of call seriousness and is included because, during training, potential responders are told to call EMS in any situation serious enough to warrant the use of supplemental oxygen. It is therefore expected that the correlation between O2 and EMS would be very high, yet this is not the case. The Pearson correlation coefficient is 0.48 indicating a fair amount, but not complete, correlation; deviations can be explained by the finding that there is substantial administration of oxygen to less serious patients.

Variable Set			
Variable	Min	Max	Description
Year	2010	2014	Year of incident
O2	0	1	Supplemental Oxygen Used No=0, Yes=1
RR1	8	36	Respiration Rate recorded for Vitals Check
EMS	0	1	Emergency medical services called No=0, Yes=1

Table 2.1: Variables used in analysis of supplemental oxygen decisions.

2.2.1 Missing vital signs

In the original set of observations, 315 of the available 555 observations (57%) are missing information on patient vital signs. For patients with minor injuries vital signs are often not collected by MFRs. Typically this is because a vital sign check would seem invasive when all that is required is very basic first aid. This is not the only reason for missing vital signs, however. The patient might have refused, or for a very small proportion of cases, the call might have been too serious to record vital signs prior to the arrival of EMS, for example. The majority are non-serious cases of minimal first-aid. If these cases with missing vital signs are excluded the sample is substantially biased towards more serious patients. Therefore the missing records could have a

substantial impact on the estimate of overall adherence to guidelines within this unit of the organization if a large proportion of cases without vital signs recorded are attributable to patients with normal range respiration rates who were correctly not administered oxygen in accordance with the guideline. In order to improve the representativeness of the sample, missing data values were imputed. For the purposes of imputation, all respiration rates associated with cases in which vital signs were not recorded were assumed to fall within normal range. Based on this assumption, missing respiration rates were imputed by taking a random draw of the normal-range respiration rates. Personal communication with the municipal organization's unit Chief in 2015 confirmed this as an acceptable representation of the vast majority cases for which respiration rates were not recorded. Although this is not the only reason for a missing record, it is the most common. Replacing the missing respiration rates with 315 random draws (with replacement) from the pool of observed normal range respiration rates increases the set of complete records to 555 observations. All analysis is undertaken using this augmented set of 555 observations. Analysis using the original data without the imputed values is contained in Appendix Section 2.A for reference. The results lead to the same overall conclusions.

Figure 2.2 provides a visual summary of the key variables used in the analysis. The first pane highlights the unbalanced nature of oxygen administration decisions; there are nearly double the number of observations for 'Do not Administer' as there are for 'Administer'. The second pane highlights the fact that most treatment given by MFRs is not serious enough to warrant a call to EMS. In the third pane the Epanechnikov kernel density of respiration rates (RR1) is presented. The location of the mean of this distribution to the right of the mode suggests that this variable is not normally distributed. A Shapiro-Wilk test confirms this result.

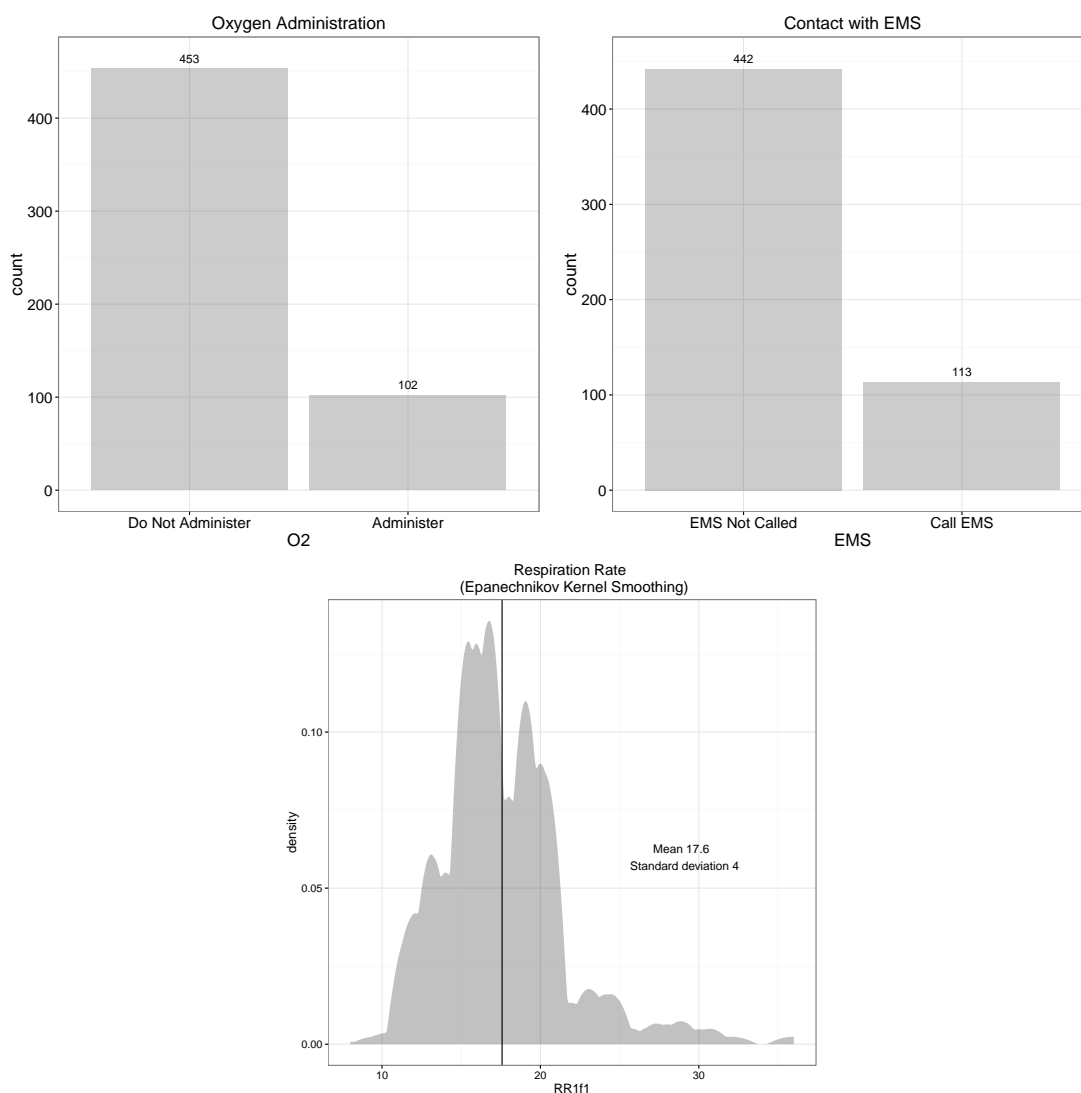


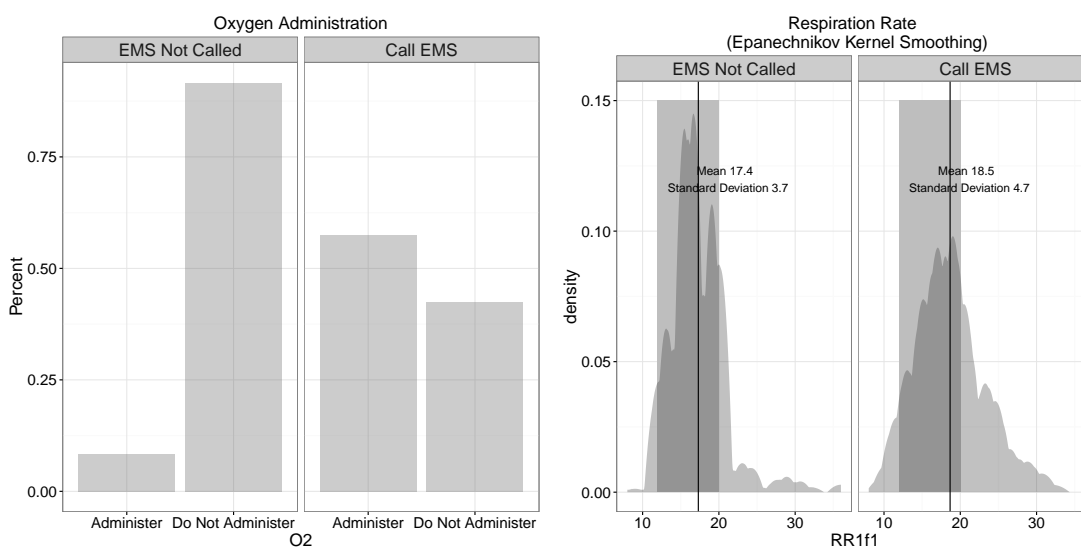
Figure 2.2: Illustrated data summary.

2.2.2 Contact with emergency services

In addition to the overall summary data reported in Figure 2.2, Figure 2.3 divides the data across categories of EMS Contact: 'EMS Not Called' and 'Call EMS'. Contact with Emergency Medical Services (EMS) is an indicator of the seriousness of an incident. A chi-square test for independence rejects the notion that Oxygen Administration, Respiration Rate and EMS calls are independent. Oxygen Administration is found to have significant correlation

with EMS contact (0.51 with a p-value of 0.0000 using the Pearson product-moment correlation), but not so much as to represent multicollinearity. MFRs are trained to respond in serious emergencies according to procedures set out by the medical director of the organization. Acceptance to the role of MFR is dependent on the ability of candidates to follow such procedures accurately and this training is reviewed frequently. Serious cases receive much attention in training, and there is little to no room for mistakes in the administration of oxygen or making contact with EMS in such cases. Failure to apply oxygen when it is necessary can have substantial adverse health consequences for patients. In examinations, failure to apply oxygen when needed disqualifies an MFR candidate. In less serious cases patients demonstrate less frequent need for oxygen or EMS assistance. A difference in adherence across serious and non-serious cases is possible if more training for serious incidents affects adherence rates when such cases are encountered in the field. Figure 2.2 demonstrated that in the field the majority of incidents (approximately 80% of cases) are non-serious. MFRs thus also acquire experience in dealing with non-serious incidents which may also influence adherence rates across levels of case seriousness. Unfortunately, the characteristics of individual MFR experience were not available in this data set and so contact with EMS serves as the only means of evaluating the effect of scene seriousness on guideline adherence. As well, patient outcomes following treatment are unavailable so the health impacts of MFR services are unknown.

Figure 2.3 presents the results by EMS contact sub-groups. The first pane makes clear that when contact with emergency services is not made, oxygen is usually not administered. The second pane highlights a skew towards normal range respiration rates when EMS Services are not contacted. In this pane the Epanechnikov kernel smoothed density of respiration rates shows a marked difference in distributions across EMS contact levels. Normal range



EMS is emergency medical services.

Figure 2.3: Illustrated data summary by EMS contact sub-group.

respiration rates are highlighted in the darkened grey box.⁵ These results inform the inclusion of EMS contact as an important marker of call seriousness, as illustrated by the difference in respiration rates, and also as an influence upon oxygen administration decisions, as shown by the relative percentages of administration decisions.

2.3 Method stage 1

The first stage of assessment involves simply sorting each recorded oxygen administration decision according to whether or not these decisions adhere or do not adhere to the guideline, based on the respiration rate associated with the decision. A classification matrix provides a concise means of summarizing the results. Each cell of a classification matrix represents the number of observations which satisfy the conditions specified in the row and column. The basic format is illustrated in Table 2.2.

Table 2.2 labels the totals in each cell. True negatives for example, are data

⁵Figure 2.A.2 presents the result without the imputed values. The difference remains.

		Observed Decisions	
		0	1
Guideline	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Table 2.2: Classification matrix guide.

points in which no action was taken and the guideline suggested no action should be taken. These decisions are thus adherent to the guideline. Similarly, decisions in the cell row labelled 1 and column labelled 1 indicate decisions in which action was taken and the guideline suggested that action should be taken. These decisions thus also adhere to the guideline. In terms of oxygen administration decisions the possible categorizations entail:

True Positive: The guideline recommends oxygen administration and oxygen was administered.

True Negative: The guideline recommends no oxygen administration and oxygen was not administered.

False Negative: The guideline recommends oxygen administration and oxygen was not administered ('under administration').

False Positive: The guideline recommends no oxygen administration and oxygen was administered ('over administration').

The classification matrix approach thus provides information, not only on adherence to guidelines, but on non-adherence as well. Estimates of over and under provision of care with reference to a guideline can be made easily.

2.3.1 Summary measures of correct classification

Table 2.3 presents a set of measures which can be used to summarize the classification matrix. Because the classification matrix is unit-free these mea-

Measure	Formula
True Positive Rate	$TPR = \frac{TP}{TP+FN}$
True Negative Rate	$TNR = \frac{TN}{TN+FP}$
False Positive Rate	$FPR = \frac{FP}{TN+FP}$
Correct Classification Ratio	$CCR = \frac{TP+TN}{Total}$
Balanced Correct Classification Ratio	$bCCR = 0.5(TPR + TNR)$
Area Under the Curve	$AUC = \int_{-\infty}^{\infty} TPR(t)FPR'(t)dt$
Cohen's kappa	$\kappa = \frac{p_0 - p_e}{1 - p_e}$

TN= True Negatives, TP = True Positives, FN= False Negatives, FP = False Positives, and Total is the sum of all entries in the classification matrix: TP+TN+FN+FP.

t is the threshold used for sorting observations.

$$p_e = \frac{(TN+FN)}{Total} * \frac{(TN+FP)}{Total} + \frac{(FP+TP)}{Total} * \frac{(FN+TP)}{Total}$$

$$p_0 = \frac{TP+TN}{N}$$

Table 2.3: Classification matrix summary measures.

measures are generally comparable across applications. It is important to note, however, that not all the measures listed in Table 2.3 are equally suitable for all data sets. For example, the Correct Classification Ratio (CCR) is the proportion of adherent decisions in the data and is therefore a basic summary measure of adherence. The CCR does not account for the fact that the more frequent outcome has a greater probability of adhering to the guideline simply by chance. Straube and Krell (2014) also note that this measure is biased when the data are unbalanced, which, as noted in Section 2.2 is the case for the illustrative application.

Four metrics are suggested by Straube and Krell (2014) to improve the informative capacity of classification matrix summarization on the basis of being insensitive to class imbalance: The Balanced Correct Classification Ratio

(bCCR), the Geometric Mean (G-mean), Area Under the Receiver Operator Characteristics Curve (AUC) and d-prime. The bCCR and AUC metrics are applied here because these are both simple to understand and insensitive to class-imbalance. The bCCR is related to the CRR, and improves upon the CCR by taking into account class imbalance. If the the sample of data used were balanced the bCCR with be equivalent to the CCR. The AUC is another measure of the degree of match between observed outcomes and the guideline. The AUC is a proportional measure which ranges from 0.5 to 1.00 and is commonly used to evaluate the match between predictions and observations over a broad a range of scientific settings, thus making this measure accessible to a variety audiences. An approximate classification of the level of match for AUC values is presented in Table 2.4.

AUC Value	Performance
0.5 - less than 0.6	Fail
0.6- less than 0.7	Poor
0.7 - less than 0.8	Fair
0.8 - less than 0.9	Good
0.9 - 1.00	Excellent

Source: Tape (2015).

AUC is area under the receiver operator characteristics curve.

Table 2.4: Classification of AUC values.

At this stage of analysis outcomes are binary (0 for ‘Do Not Administer’ decisions and 1 for ‘Administer’ decisions) so the threshold (t) is irrelevant, but Section 2.6 will discuss the AUC in greater detail, making use of this measure with relevant threshold values. Jeni, Cohn, and De La Torre (2013) show that the AUC is insensitive to skewness, where $skewness = \frac{FN+TP}{TN+FP}$, but that this measure can mask poor performance of a classifier. Fortunately, such masking is not an issue at this stage of analysis due to the binary nature of the out-

comes.⁶

Cohen's κ (Cohen, 1960) is used to assess the amount of agreement between the guideline and the observations beyond that which would be achieved by chance. Jeni et al. (2013) show that Cohen's κ is sensitive to both the degree of skewness and rate of misclassification in simulation experiments. However, for rates of skewness between 0.1 and 10 and a misclassification rate of 1%, the accuracy of Cohen's κ is near 95%, so this measure will be quite accurate for many data sets. While no single metric dominates in terms of providing information, Cohen's κ is important for summarizing the rate of adherence to a guideline net of chance.

2.4 Application: Stage 1 adherence results

This section reports the Stage 1 results of the standardized analysis of the guideline evaluation framework applied to the data described in Section 2.2. As previously mentioned, these data showed substantive differences across sub-groups of EMS contact, therefore the analyses are displayed for each sub-group.

Table 2.5 presents the classification matrix from the sub-group defined by non-contact with EMS. These cases required less serious first-aid treatment. Non-applications of oxygen feature prominently in this sub-group, with a skewness measure of 0.09. This level of skewness implies that the use of skew-insensitive classification matrix summary metrics are appropriate.

Non-adherent decisions represent a very small share of the total decisions for this sub-group and there is virtually no difference in the type of non-adherent

⁶A classifier is a model which predicts outcomes. These measures are often used in evaluating model performance, i.e. the ability of a predictive model which generates values between 0 and 1 to match with the observed outcomes which are 0 or 1. This stage of the analysis simply evaluates the match of the observed outcomes, which are 0's and 1's and thus do not require a threshold for sorting, with the guideline recommendations (0 or 1).

decisions across sub-groups. The equal distribution of non-adherent decisions across categories of under- and over- administration provides no evidence to support the claim that MFRs have a tendency towards administering oxygen to non-emergency patients as a comfort measure any more than they have a tendency towards neglecting to administer oxygen to patients who demonstrate need for it based on respiration rate in non-emergency situations.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	376	30	406
	Administered	29	7	36
	Total	405	37	442

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 2.5: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called.

Table 2.6 tells a different story for the sub-group defined by EMS contact.

This sub-sample is more balanced, with a skewness measure of just 1.35 which indicates a very slightly larger share of 'Administer' decisions than 'Do Not Administer'.⁷ Non-adherent decisions in this sub-group represent a larger absolute share of decisions than in the non EMS contact sub-group (49% vs 13%), and this difference is significant using a two sample test for equality of proportions with continuity correction (p-value of 0). As well, the type of non-adherence exhibited within this sub-group is significantly biased towards administering oxygen when the guideline recommends against it (p-value of 0, using a two-sample test for equality of proportions with continuity correction). This result combined with the overall greater proportion of administration decisions made in this sub-group suggests that when faced with

⁷Skewness values in the range of [0.1-10] are symmetric around 1. Thus, a skewness value of 9.14 would be an equivalent degree of skewness towards 'Administer' decisions as the skewness towards 'Do Not Administer' decisions observed in the sub-group defined by non-contact with EMS.

a serious scene MFRS tend towards administering oxygen regardless of the guideline.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	37	44	81
	Administered	11	21	32
	Total	48	65	113

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 2.6: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called.

The classification matrix summary measures presented in Table 2.7 suggest that the guideline is not well followed. The substantial impact of skewness is illustrated in the EMS-not-called sub-group as the CCR drops by 36% moving to the skew-insensitive bCCR. Across sub-groups the skew-insensitive metrics bCCR and AUC are virtually identical. The AUC results suggest failure to adhere with the guideline. Cohen's κ suggests that only 12% of decisions above and beyond chance are adherent to the guideline in the EMS-not-called sub-group, and only 9% of decisions are adherent to the guideline in the EMS-called sub-group. In the EMS-not-called sub-group the level of skewness suggests that Cohen's κ may even be slightly overstated.

Measure	EMS Not Called	EMS Called
TPR	18.92%	32.31%
TNR	92.84%	77.08%
FPR	7.16%	22.92%
CCR	86.65%	51.33%
bCCR	55.88%	54.7%
AUC	0.56	0.55
Cohen's κ	11.9%	8.62%

Table 2.7: Classification matrix summary measures by EMS contact sub-group.

2.4.1 The economic consequences of non-adherence

Guidelines are expected to encourage the appropriate use of medical resources. This section estimates the annual cost to deliver supplementary oxygen to patients and the potential savings from greater adherence to the administration guideline. The organization reports annual oxygen costs of \$400 per year. Over 2010-2014 a total of 898 patients were treated. Approximately 30% of these cases were children under the age of 18 (271 cases), which follow a different oxygen administration guideline. Attributing 70% of oxygen costs to adults implies an annual cost of \$280. Costs attributable to each sub-group are obtained by simply dividing the costs in proportion to the number of treated patients in each sub-group.

Table 2.8 reports the estimated costs of oxygen per patient treated with oxygen. Training in correct oxygen delivery is a substantial part of MFR training and oxygen delivery equipment represents approximately half of the total amount of equipment required to be carried by MFRs while on duty. Table 2.8, however, reflects that despite the substantial training and equipment devoted to oxygen administration, overall oxygen use is infrequent. In total the sample contains 37 instances of oxygen use when EMS was contacted and 65 total cases when EMS was not contacted, as reported in Tables 2.6 and 2.5. The mean annual number of patients treated is estimated by taking the mean of the number of patients in each year from 2010-2014. Only patients treated with oxygen incur costs attributable to oxygen,⁸ so costs per patient are simply total costs per year divided by the average number of treated patients per year. The table reflects that oxygen was administered to patients both when EMS was called and when EMS was not called. Since responders may behave differently when faced with more serious situations, such as those requiring contact with EMS, these sub-groups are presented in separate columns.

⁸There is insufficient information to justify different costs across levels of adherence.

Measure	EMS Not Called	EMS Called
All Cases		
Treated Patients	7.4	13
Total Costs	\$ 101.57	\$ 178.43
Cost/patient	\$ 13.73	\$ 13.73
Guideline Non-Adherent Cases		
Over-Administration		
Treated Patients	6	8.8
Unjustified Costs	\$ 82.38	\$ 120.82
Under-Administration		
Untreated Patients	5.8	2.2
Unjustified Savings	(\$ 79.63)	(\$ 30.21)
Perfect Guideline Adherence		
Treated Patients	7.2	6.4
Total Cost	\$ 98.86	\$ 87.87
% Savings under Perfect Adherence	3%	51%

Table 2.8: Annual cost of oxygen delivery by EMS contact sub-group.

To estimate costs under perfect adherence, the mean number of patients per year is simply multiplied by the cost per treated patient to obtain a figure for the average cost all treated patients. The cost of over-treatment (over-administration) is then deducted from the total cost figure for each sub-group and under-treatment added. The percentage of actual cost that will be saved under perfect adherence is equal to 100 times the ratio of the difference between the actual cost and the cost with perfect adherence to the actual cost. It cost donors and community sponsors \$13.73 to administer oxygen to a patient over the five years 2010 through 2014. If adherence to the guidelines was perfect a 3% reduction in expenditures would have been realized for non-serious cases, and a 51% reduction in expenditures would have been realized for serious cases. Bearing in mind that only 20% of all patient care reports reviewed were serious enough to warrant contact with EMS (21% per year on average), efforts to encourage greater adherence may be better directed at other activities within the organization. In addition to representing only

a minor level of costs, a large proportion of non-adherent decisions occur in serious situations and constitute over-administration. In serious emergencies supplemental oxygen has potentially life-saving benefits and has a low risk of adverse medical events associated with administration. The benefits of such 'better safe than sorry' over-administration likely outweigh the costs for the organization. The continuous training to recognize serious emergencies approach rather than strict guideline enforcement appears to serve emergency patients who require oxygen rather well, while carrying minimal negative impacts on patients who do not need oxygen. In this situation advocating for strict guideline adherence may have a negative impact on oxygen administration practices. This is because a strict policy of guideline adherence could very well crowd-out the MFR's focus on accurate scene assessment and patient well-being which informs administration of oxygen in a serious situation and redirecting attention towards recalling and implementing guidelines. The reaction might be seen as a type of 'crowding-out' of focus; Frey and Jegen (2001) give examples of the empirical relevance of such effects. The potential for stricter adherence strategies to induce unanticipated negative reactions warrants careful examination of the behavioural context in which a guideline is situated and should ideally be carried out prior to implementation.

2.5 Method stage 2

The classification matrix approach is a useful first stage in evaluation, however many studies take a regression approach to evaluating guideline adherence. When there are more than 2 or 3 covariates the classification matrix approach can become unwieldy, as a set of matrices must be computed for each level of each covariate. Stage 2 of the analysis presents a state-of-the-art regression framework. This framework removes the need to select an appropri-

ate parametric form for estimation, eliminating errors arising due to model misspecification. The framework also has variable selection embedded, as covariates which do not influence the outcome are eliminated from the regression. Section 2.6.1 compares the performance of the regression framework proposed to various alternatives using the AUC as a metric for comparison. In economics, applications involving evaluation of binary outcome data, such as the 'Administer' and 'Do Not Administer' outcomes in the example, rely heavily upon parametric smoothing approaches such as probit, logit, or linear regression, sometimes with higher order terms. This preference is reflected in the health economics literature. An EconLit search of 'Guideline*' and 'Medical' returned 115 articles. After sorting for guideline adherence 30 were retained and reviewed. The main analysis styles applied were survival analysis, probit regression, logit regression and OLS or variants thereof (fixed effects or random effects). The selection of a particular regression model is often dictated by the type of data encountered, and the guideline itself. As will be illustrated in Section 2.6.1 below, these regression models can inaccurately identify distinct patterns within a set of observations which appear to match with the profile of guideline adherent decisions. Because of this, I propose a regression approach that applies nonparametric conditional density estimation to assess the pattern of observed treatment decisions. This approach generates a profile of estimated probabilities of administering oxygen over the range of respiration rates for each level of EMS contact. The profile of estimates may take on any form, and so may or may not follow a pattern reflective of the guideline itself. The guideline need not even be known at the outset of applying this regression strategy.

The guideline instructs administration of oxygen to adult patients when respiration rates are less than 12 breaths per minute, no administration of oxygen when respiration rates are in the normal range of 12 to 20 breaths per

minute, and administration when respiration rates exceed 20 breaths per minute. Plotting the guideline over a range of respiration rates results in a solid line at 'Administer' with a single downward step the 'Do Not Administer' at a respiration rate of 12, and a single step upwards, back to 'Administer' at a respiration rate of 20, as was shown in Figure 2.1. If the observations reflect this guideline it is expected that the estimates generated by the regression will show a region of concentrated change in the negative direction at 12 breaths per minute and a region of concentrated change in the positive direction at 20 breaths per minute. These changes in the estimates are summarized by the gradients (instantaneous rates of change). The minimum and maximum valued gradients of the estimates are used to identify the respiration rates where the most concentrated changes occur in the profile of estimates for each sub-group, and over all encountered respiration rates. The identified respiration rates form 'candidate steps' which can then be compared to the guideline 'steps' of 12 and 20. Bootstrapped confidence intervals are used for this comparison.

Nonparametric conditional density estimation makes as few distributional assumptions about the data as possible. In a regression context, this means that issues of model misspecification arising due to incorrect parametric assumptions about errors are avoided. Because no form is specified, it also means that the patterns unmasked by such estimation may serve as the basis for a hypothesis test. Estimates resulting from nonparametric estimation techniques may be better able to suggest evidence for or against guideline adherence than parametric alternatives alone. Three other approaches to evaluation are illustrated for comparison: an Empirical approach which plots simple proportions, a parametric regression which fits a quadratic specification of the relationship between respiration rate, call seriousness and oxygen administration; and a search approach which simply chooses the guideline of best

fit based on maximization of the AUC.

Nonparametric conditional density estimation is undertaken in the statistical environment R (R Core Team, 2015b) using the *np* package (Hayfield and Racine, 2008). All that is required is to first compute optimal bandwidths and secondly to generate the estimated values. All interactions between covariates are automatically accounted for and irrelevant covariates are smoothed out asymptotically. It is important to note that this method does not deliver scalar coefficients as parametric regression does because the estimates are not confined to taking on a particular form (such as a linear form in linear regression). This means that statistical tests on parameter values (i.e. tests on coefficients) are generally out of the question and motivates the use of ‘candidate steps’ described above.

Smoothing nonparametrically relies on the observations rather than a pre-specified functional form for the resulting estimates. An equation with respiration rate (*RR1*) and contact with emergency medical services (*EMS*) entering as explanatory variables defines the relationship and fitting is done via the estimation of optimal bandwidths for kernel conditional density estimation. All interaction effects are automatically incorporated. Nonparametric conditional density estimation was described first by Stone (1977) and more recently described by Hall, Racine, and Li (2004). Appendix 2.B provides greater detail of the estimation strategy. All that is required for implementation is to enter the formula:

$$O2 \sim RR1 + EMS, \quad (2.1)$$

in the R computer package (R Core Team, 2015b), where *O2* is the binary outcome of the oxygen administration decision made by an individual MFR. This formula is used first in the determination of the optimal bandwidth sizes.

The routine which determines the bandwidth automatically accounts for interactions between *RR1* and *EMS*. Having computed the bandwidths, the estimates and gradients are then generated over the range of values encountered in the data. Once a profile of estimates is obtained, the identification of candidate steps is carried out by simply selecting the respiration rates associated with the lowest and highest gradients. The main aim of this methodology is to determine whether observations are reflective of a pre-set practice guideline. Generating a profile of estimates offers insight into the shape of the responses. Within the profile of estimates, identifying regions of greatest change in the negative and positive directions provides candidate steps which can now be compared to the guideline. For guidelines involving single steps up or down see Thomas (2016).

Comparison with the guideline proceeds by creating confidence intervals by bootstrapping. This method simply reconstructs the estimates using a re-sample of the original observations of the same size and with replacement. The process is repeated 1000 times and the 5th and 95th percentiles of the results taken as the boundaries of the 90% confidence interval. This avoids the need to construct a bootstrap sample which is consistent with the null hypothesis in order to carry out a full nonparametric hypothesis test, which can be a complex task, especially for non-standard estimators (MacKinnon, 2007). Support for a particular step matching with the guideline is offered if the guideline falls within the bounds of the confidence interval around a step.

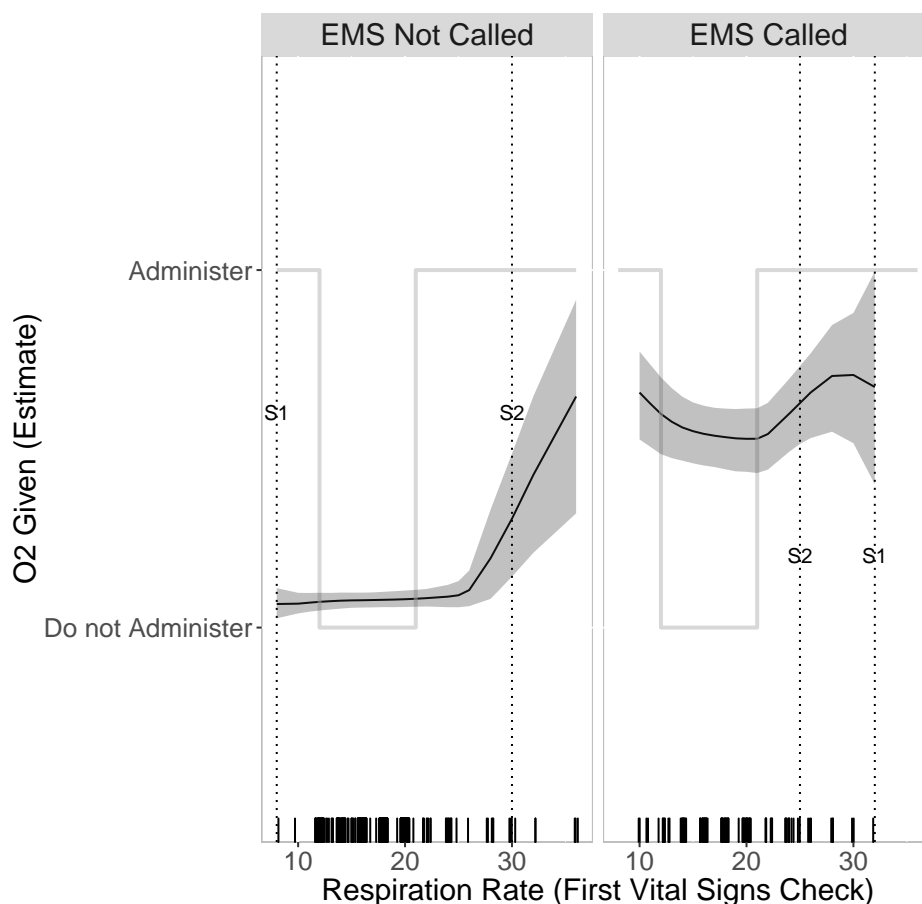
2.6 Application: Stage 2 adherence results

Figure 2.4 presents the results of the nonparametric strategy, the two 'candidate steps', and the practice guideline, along with the bootstrapped 90% confidence bounds. The results illustrate a partial effects surface, because the

nonparametric routine fits a multi-dimensional surface. The partial effects are therefore 'slices' through the three dimensional prediction object ($RR1$, EMS and the prediction \hat{O}_2) along each ridge defining whether or not EMS was called and mapped over the range of all $RR1$ recorded. All interactions are accounted for automatically and the prediction surface can take on any shape. All estimates are projected over the full range of possible values of respiration rates encountered in the data set.

Figure 2.4 makes clear that there is a difference in the probability of using supplemental oxygen depending on the sub-group defined by whether or not the situation warranted a call to EMS. The pattern which would suggest guideline adherence should have candidate Step 1 and Step 2 values in order S_1 , S_2 and near respiration rates of 12 and 20. Neither sub-group thus exhibits a pattern indicating guideline adherence. Visual inspection reveals that there is a lower overall use of oxygen in the 'EMS Not Called' sub-group. As well, there appears to be a slight indication of a U-shape in the cases where EMS was called, but the candidate steps do not identify this pattern at all. In 'EMS-called' cases the first step, the step downward to 'Do Not Administer', is located at a higher respiration rate (32) than the upward step to 'Administer', which is located at the respiration rate of 25. In 'EMS Not Called' cases the estimates slope upwards very slowly.

To test the precision of the candidate steps bootstrapped confidence intervals are constructed. Table 2.9 reports the steps, confidence intervals, guideline values and gradients for for Step 1 and Step 2 for each sub-group. The confidence intervals for Step 1 and Step 2 are nearly as wide as the data range in all cases except Step 2 in the 'EMS Not Called' sub-group, which occurs at the upper bound with greater precision than the other candidates. In both sub-groups the Step 1 and Step 2 confidence intervals overlap, indicating that two distinct steps are not identified. This is evidence against the finding of adher-



Guideline is the grey line. Candidate step 1 (S1) and step 2 (S2) are dotted lines. Bootstrapped 90 percent confidence interval is the shaded area. Bars along the x-axis indicate frequency of observations.

Figure 2.4: Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach.

ence to the guideline in both serious and non-serious cases, despite the fact that the guideline falls within the bounds of the confidence intervals in all but the Step 2 'EMS Not Called' case.

The overall result is consistent with the results of Stage 1. This Stage 2 regression approach indicates that there is little to no guideline adherence. What Stage 2 adds is the visual intuition of how decisions differ across sub-groups. The weakness of adherence to the guideline is striking under this representation. In the 'EMS Not Called' case oxygen use is less frequent, but increases

EMS	Step	Lower ci	Upper ci	Guideline	Gradient
EMS Not Called	Step 1	8	36	12	-0.0024
EMS Not Called	Step 2	30	36	20	0.0524
EMS Called	Step 1	32	32	12	-0.0377
EMS Called	Step 2	25	32	20	0.0357

Table 2.9: Candidate steps and bootstrapped 90 percent confidence intervals using the nonparametric strategy.

at very high respiration rates. In the 'EMS-called' sub-group there is an overall greater estimated use of oxygen, slightly higher at low and high respiration rates, but not in a pattern reflective of the guideline. The areas of most concentrated change in the estimates are not distinct, as demonstrated by bootstrapped confidence bounds.

2.6.1 Relative performance of stage 2

In order to assess the relative performance of the nonparametric strategy used in Stage 2, I undertook comparisons with three other strategies for condensing observations into estimates. For each strategy the profile of estimates with their associated bootstrapped confidence bands is presented. The ability of each model to correctly predict the data is evaluated using the AUC metric. Next, the location of the largest changes in the negative and positive directions are identified and 90% confidence intervals generated. These results are evaluated for the presence of distinct candidate steps and their match with the guideline.

The 'empirical' strategy simply plots the proportion of 'Administer' decisions at each respiration rate. The 'linear' strategy fits an ordinary least squares regression to the data, making use of a quadratic form and interaction effects. The 'search' method evaluates the fit of every possible specification of the guideline to the data. While the nonparametric AUC results may not always be largely different from the comparisons, the nonparametric strategy

requires the input of just one calculation and is thus a more straight-forward and efficient approach to analysis. In contrast, the search approach requires the analyst to specify each possibility and to calculate every possible outcome. The estimates also serve as a visual test for the presence of the guideline.

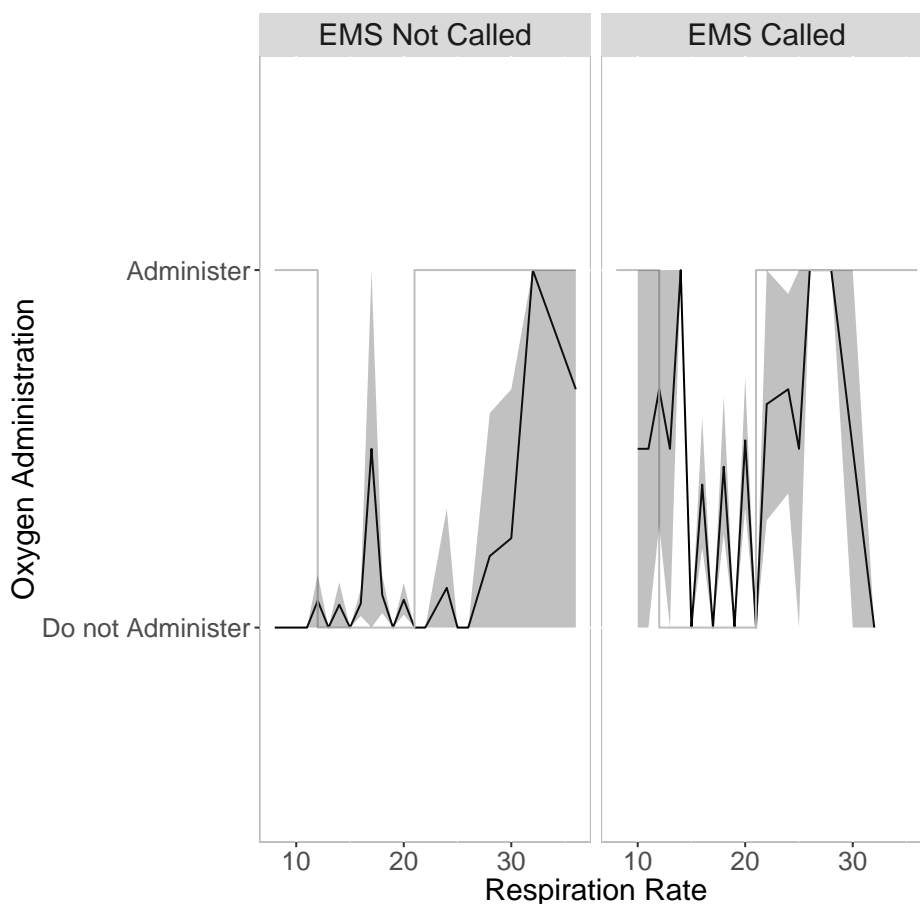
The strategies applied here can be thought of as approaches to smoothing observations. Smoothing attempts to overcome the inherent noisiness of data collected in the field. For example, the empirical strategy produces a single estimate for each level of respiration rate in each sub-group by plotting the proportion of cases in which oxygen was used over the recorded respiration rates. While some support for the reflection of a guideline is visible in the profile of estimates, which are noisy, greater smoothing could make the results clearer. For single step patterns, the main models encountered in the literature were Probit and Logit which are smooth approximations to a heavy-side step function. In this case both a step down and a step up are expected so instead of a step function an approximation to a boxcar function is more appropriate and is done here by including higher order terms in a linear regression framework. This results in a U-shaped profile of estimates over respiration rate. A third approach considers each possible specification of the guideline, choosing the specification with the best fit according to the greatest AUC. The nonparametric strategy outlined in Section 2.5 smooths, but is also responsive enough to allow for sharp changes akin to the discrete changes defining the steps downward and upward within the estimates if the data reflect such a pattern. The main point is that such a pattern need not be defined at the outset of the analysis.

2.6.1.1 Empirical strategy

As already mentioned, the empirical strategy simply plots the proportion of ‘Administer’ decisions for each sub-group over the range of respiration rates. The result is a noisy profile of estimates. The selection of candidate steps proceeds as outlined in Section 2.5. In cases where multiple minimum and maximum gradient values are encountered the median value respiration rate is reported. Table 2.10 reports the candidate steps and confidence intervals. All candidates except Step 2 in the ‘EMS Not Called’ sub-group fall within the 90% confidence interval. The intervals of Step 1 and Step 2 overlap in both sub-groups offering evidence against two uniquely defined steps. Figure 2.5 illustrates the result. The profiles of estimates are very noisy, masking any potential steps.

EMS	Step	Lower ci	Upper ci	Guideline	Gradient	
EMS Not Called	Step 1	18.0	18	36	12	-0.4091
EMS Not Called	Step 2	32.0	10	26	20	0.7500
EMS Called	Step 1	15.0	11	15	12	-1.0000
EMS Called	Step 2	22.0	10	26	20	0.6250

Table 2.10: Candidate steps of the empirical strategy and bootstrapped 90 percent confidence intervals.



Proportion of 'Administer' outcomes is solid black line and the guideline is the grey line in each panel. Bootstrapped 90 percent confidence interval is the shaded area.

Figure 2.5: Predicted administration of oxygen using the Empirical approach by respiration rate and EMS contact sub-group.

2.6.1.2 Linear approach: regression with higher order terms

One approach to obtaining smoother profiles of estimates than those of the Empirical approach is to specify a linear model with higher order terms which result in 'U' shaped profiles of estimates in each sub-group. Here fitting is carried out using a simple Ordinary Least Squares approach with the form ⁹

$$\hat{O}_2 = \alpha + \beta_1 EMS + \beta_2 RR1 + \beta_3 RR1^2 + \beta_4 RR1 * EMS, \quad (2.2)$$

⁹Since this is a binary outcome a Probit model with an index function based on the quadratic specification of equation 2.2 is also appropriate. The results are virtually identical and omitted here for simplicity.

where α represents an intercept, EMS is an indicator of call seriousness equal to 1 if EMS services were contacted and 0 otherwise, and RR1 and RR1² the recorded respiration rate and respiration rate squared and RR1 * EMS the interaction of respiration rate and contact with EMS.¹⁰ The regression estimates, presented in Table 2.11, suggest that all covariates except the interaction term are significant at the 2% level or less. A Ramsey RESET test for correct specification fails to reject the null of a correctly specified model (p-value 0.7).

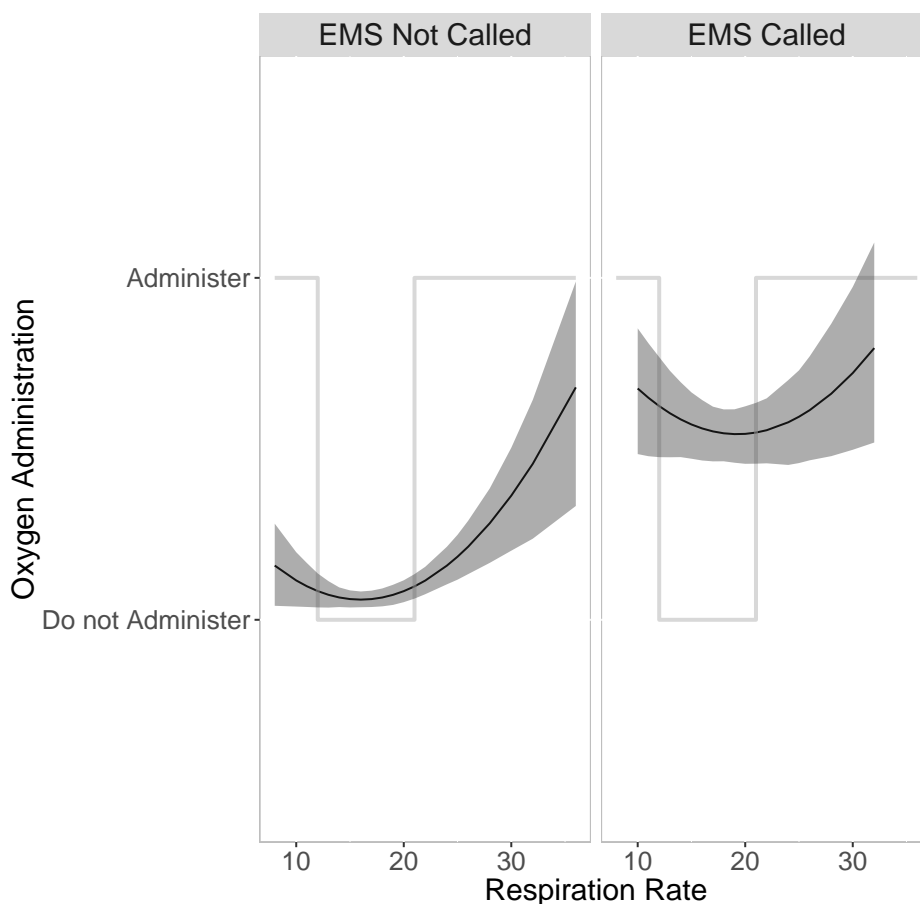
Figure 2.6 plots the estimates from fitting Equation 2.2. Two clear u-shapes are apparent in each of the levels of call seriousness. As well, there is a marked difference in the proportion of oxygen use across levels of call seriousness: more serious calls have a 47% higher chance of oxygen administration. Overall, the unadjusted R^2 value of this regression is 0.29, suggesting that variation in respiration rate and call seriousness contributes to just 29% of the variation in the decision to apply oxygen.

	Estimate	Std.Error	t.value	Pr(> t)
Intercept	0.4569	0.19	2.35	0.02
RR1	-0.0497	0.02	-2.57	0.01
RR1 ²	0.0016	0.00	3.33	0.00
EMS=1	0.6628	0.15	4.41	0.00
RR1*EMS=1	-0.0101	0.01	-1.27	0.21

Table 2.11: Linear regression estimates with interaction and squared terms.

The drawback of this type of line fitting is that while there are two 'U' patterns they are not indicative of the discrete changes suggested by the guideline. Two discrete steps need to be identified in order to effectively compare these results to the guideline. Using the same method outlined in Section 2.5 leads to very precise estimates of candidate steps, identifying the boundary

¹⁰The same specification with the addition of a cubic term for *RR1* was also run, none of the coefficients associated with *RR1* were significant while the adjusted coefficients of determination were the same under both specifications (0.28 with added cubic and squared terms, and 0.28 with only the added squared term). The additional squared and cubic terms are highly collinear and the model is over-fitted.



Estimated outcome is solid black line and guideline is grey line in each panel. Bootstrapped 90 percent confidence interval is the shaded area. Estimates include interaction and squared terms.

Figure 2.6: Predicted administration of oxygen using the linear approach by respiration rate and EMS contact sub-group.

values as the candidate steps in 90% of the cases. But these boundary values are identified as candidates due to the specification of the model, and so may or may not be reflective of the raw observations. Under the linear model the largest gradients always occur at the upper and lower bounds regardless of the data at hand. In this case these bounds do not contain the guideline values of 12 for Step 1 or 20 for Step 2. However, for a guideline at the lower and upper bounds the guideline will always match the steps identified with the linear method purely by coincidence. As well, the size of the gradients at the

candidate steps indicate that the size of the change in the profile of estimates is not very large. Essentially the method precisely identifies weak candidates. See Table 2.12 for the candidate steps, confidence intervals and gradients.

EMS	Step	Lower ci	Upper ci	Guideline	Gradient	
EMS Not Called	Step 1	8.0	8	8	12	-0.0249
EMS Not Called	Step 2	36.0	36	36	20	0.0621
EMS Called	Step 1	10.0	10	10	12	0.0333
EMS Called	Step 2	32.0	32	32	20	0.0271

Table 2.12: Candidate steps of the linear strategy and bootstrapped 90 percent confidence intervals.

2.6.1.3 Search approach: the best fit guideline

As the extreme alternative to the linear regression with higher order terms the contrasting strategy is to assume that the estimates in fact take on the same profile as the guideline, but with unknown locations of the steps. This form is pre-specified by identifying each combination of respiration rates in the set of all possible combinations of respiration rates, with each set representing a candidate guideline with two steps. For each sub-group, calculating the AUC for each candidate against the observations, the maximum AUC defines the best candidate, which is a combination of a Step 1, downwards, and a Step 2, upwards. In this case there are 210 unique combinations with Step 1 occurring before Step 2, and 1 maximum AUC is identified. Figure 2.7 illustrates the result for each sub-group.

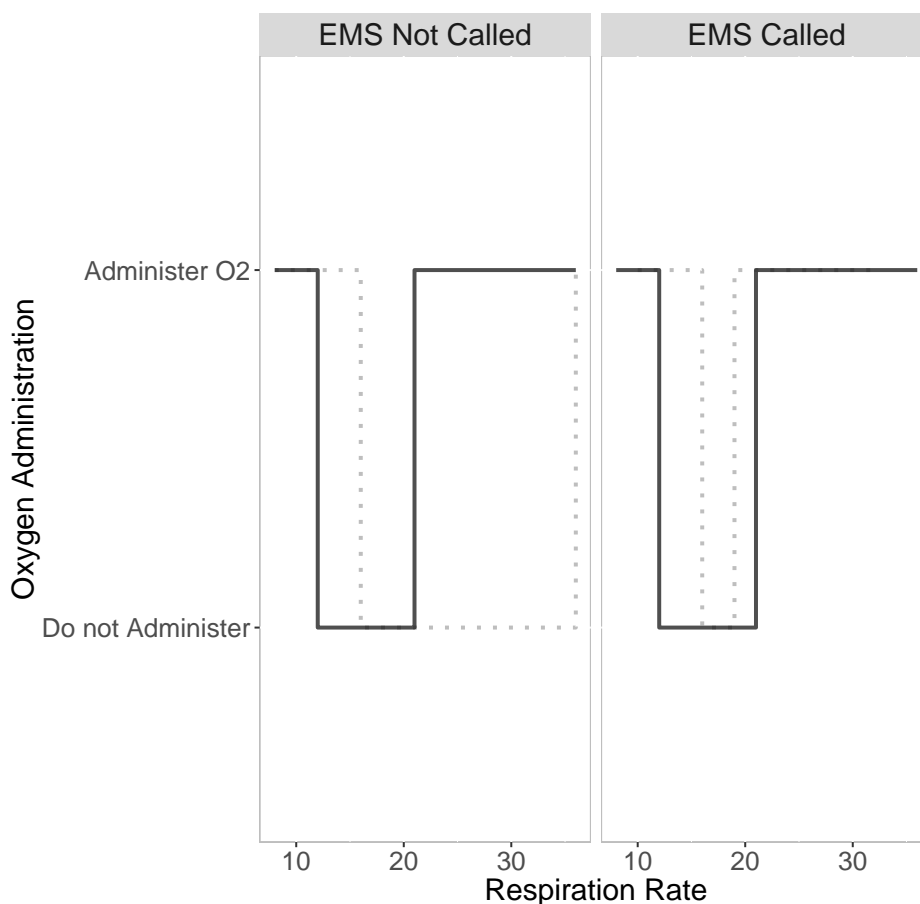
This method is a simple way to search for a candidate, but relies on the observations only in the sense of finding the best fit relative to other candidates. It is possible that the observations do not suggest such a pattern at all, in which case searching for the guideline by specifying the guideline is counter productive in determining what happened in the field.

Table 2.13 shows the results. Sometimes the bootstrap samples used to gen-

erate confidence intervals result in the identification of multiple maxima, in such cases the median value of the identified steps is reported (e.g. the median of the Step 1 values associated with the same maximal AUC value was taken as the result of a particular bootstrap sample.). In the same manner as all other bootstrap intervals the 5th and 95th percentiles of all 1000 of the resulting bootstrapped steps form the lower and upper confidence interval values for each of the steps calculated from the original data set. When EMS was called the candidates are within the bounds of the confidence intervals. When EMS was not called the steps occur outside the bounds of the confidence regions, suggesting that this calculation method did not identify significant candidates. In both levels of contact with EMS the confidence intervals overlap, suggesting that the steps are so imprecise that they do not reflect two distinct steps. This strategy hinges on specifying the distinctive step pattern which the guideline predicts, but does not indicate the true profile of estimates, only the optimal placement of the candidate guideline relative to all possible placements.

EMS	Step	Lower ci	Upper ci	Guideline	Gradient	
EMS Not Called	Step 1	15.0	16	18	12	-1.0000
EMS Not Called	Step 2	36.0	18	25	20	1.0000
EMS Called	Step 1	14.0	8	18	12	-1.0000
EMS Called	Step 2	19.0	16	30	20	1.0000

Table 2.13: Candidate steps of the search strategy and bootstrapped 90 percent confidence intervals.



Back solid lines represent the guideline. Grey dotted lines represent the optimal AUC outcome.

Figure 2.7: Optimal guideline using the search approach by respiration rate and EMS contact sub-group.

2.6.1.4 Comparison

In order to assess the performance of each approach, the goodness of fit of the profile of estimates with the observations is considered using the area under the receiver operator characteristics curve (AUC). The AUC is chosen because it is largely insensitive to imbalance in the proportion of positive to negative outcomes in the data ¹¹ and easy to calculate.

In order to assess the match of the predictions, which range from $[0,1]$, with the observations, which are binary (either a 0 or a 1), a threshold is used to

¹¹See Figure 1 in Jeni et al. (2013).

classify the predictions into 0 and 1 categories. Once categorized, the True Positive Rate (TPR) and False Positive Rate (FPR) are calculated. The process is repeated for each possible threshold value and the results plotted, forming a Receiver Operator Characteristics Curve (ROC) which describes the ability of the predictions to accurately match the observations at hand. Strategies which do not match the observations result in a 45° line extending from the origin while approaches with a high degree of accuracy curve towards the upper left of the plot. The area under the ROC curve (AUC) takes on a value between 0.5 and 1, with higher values indicating better performance.

Approach	AUC
Empirical	84.51
Linear	79.60
Search: EMS Called	54.78
Search: EMS Not Called	55.72
Nonparametric	83.14

Table 2.14: AUC values of each approach

Table 2.14 presents the results of each approach, calculated with the ‘pROC’ package in R (Robin et al., 2011). The empirical and nonparametric perform similarly, followed by the linear and search approach.¹² In particular, according to Tape (2015), the search approach fails to match the observations. The Nonparametric approach exhibits the best performance of any of the smoothing approaches (i.e. excluding the Empirical approach) and is a good fit with the observations. Two insights are gained from assessing the AUC values in this way. The first is that even the best fit ‘rectangular-u’ fails to describe the observations. The second is that, among the smoothing approaches, the Nonparametric approach describes the observations the best, and required no input from the analyst about the ‘shape’ of the estimates, unlike the search and

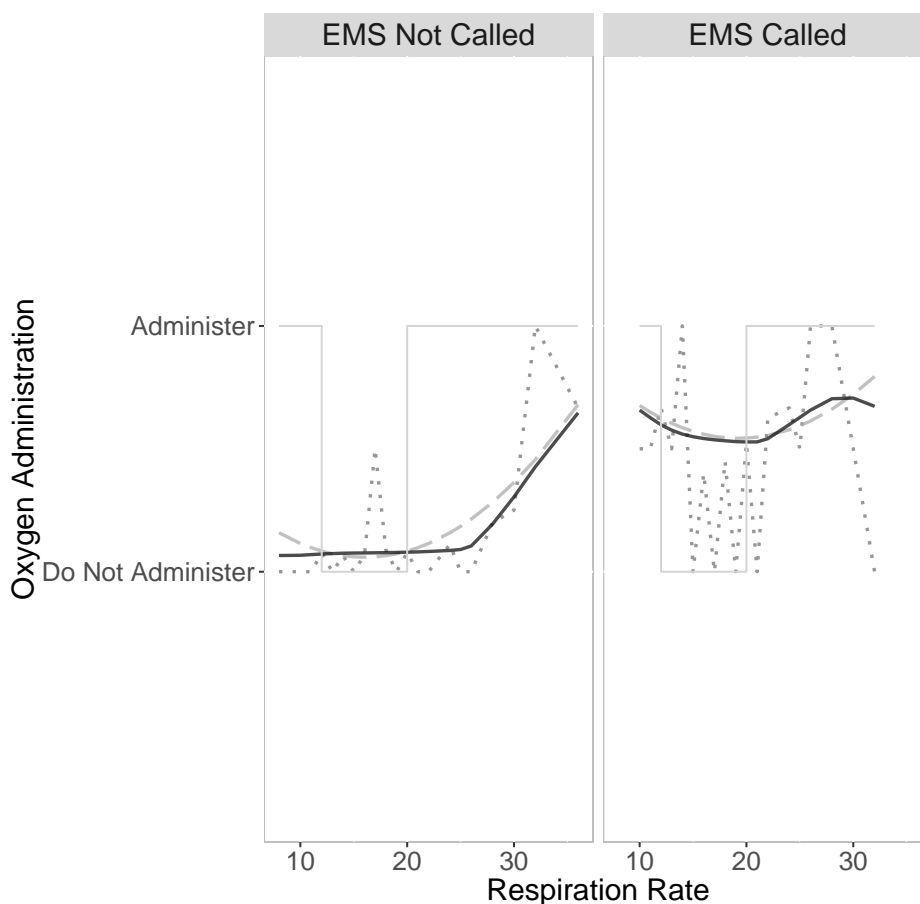
¹²The AUC is calculated over the entire range of estimates for the linear and nonparametric strategies. The AUC for each sub-group is presented for the search method due to the manner in which this strategy is formed: by choosing the guideline which maximizes the AUC in each sub-group.

linear methods. This makes nonparametric conditional density estimation an excellent approach for exploratory data analysis, and especially useful for assessing the behaviour of observations independently of presumptions about guidelines. The results of a bootstrapping test of the differences between the AUCs for the Empirical, Linear and Nonparametric strategies, presented in Table 2.15 confirms this result. The nonparametric strategy has a significantly greater AUC than the linear strategy but is not significantly different from the Empirical strategy at the 10% level of significance.

Approach	Empirical	Linear
Empirical		
Linear	0.00295	
Nonparametric	0.24632	0.06646

Table 2.15: P-values of bootstrap tests of differences in areas under ROC curves.

Layering the estimates of each approach, the result in Figure 2.8 is visually striking. While the linear approach fails to reject a null of misspecification, the better performing nonparametric estimates give no indication of a u-shape in the case of non-serious cases and only a slight u shape for serious cases. Finally, using the information in Tables 2.9, 2.10, 2.12 and 2.13, a visual summary of the identified candidate steps can be constructed. Figure 2.9 presents the two candidate steps for each strategy along with the confidence intervals associated with each step as a shaded rectangular area. The height of each confidence area is scaled to the height of the gradient associated with the candidate step. The figure thus summarizes both the precision and the degree of change associated with each candidate step. Confidence areas which are narrower indicate more precise estimates, while areas which are wider indicate less precision. Confidence areas which are taller indicate stronger gradients while shorter areas indicate smoother declines in the profile of estimates. In all cases distinct confidence areas between Step 1 and Step 2 which each con-



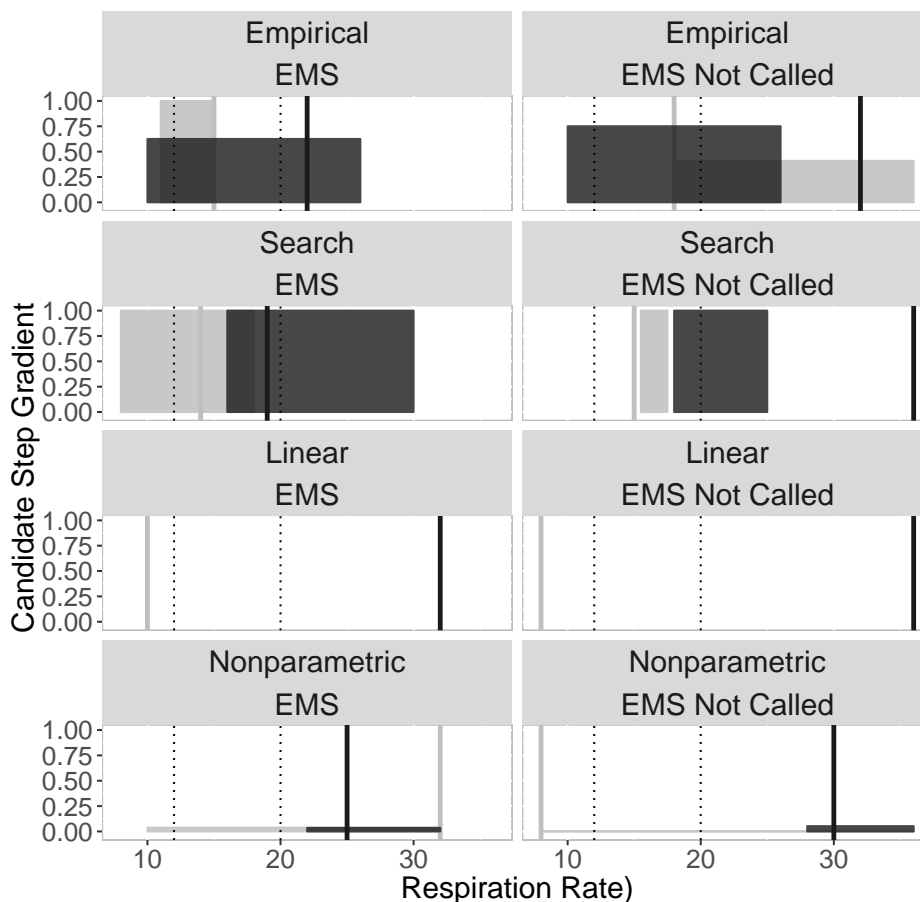
Guideline is the solid grey line in each panel. Empirical estimates are dotted grey lines, linear estimates are the dashed grey lines and nonparametric estimates are the solid black lines.

Figure 2.8: Predicted administration of oxygen for smooth approaches by respiration rate and EMS contact sub-group.

tain the guideline steps indicate support for the guideline.

Taking a closer look at the values of the gradients provides insight about the strength of the candidate steps. Gradients capture the change in the outcome at each evaluation point. Gradients for binary data can range from $[-1,1]$ since the largest steps possible are from 0 to 1 ('Do Not Administer' to 'Administer') or 1 to 0 ('Administer' to 'Do Not Administer'). The ideal pattern for a match with the guideline would be defined by a gradient of -1 for Step 1 and 1 for Step 2 which is what is observed for the search approach due to the

complete specification of the estimates. The search approach demonstrated the worst fit to the observations, however, and so the finding of such strong gradients is spurious. The empirical steps are next largest, and are indicative of the jagged path of the estimates, these estimates were the most sensitive to noise. The linear gradients are very small, due to the specification of the form of the estimates and the least sensitive to noise. The linear approach indicated two well-defined 'U' patterns in the estimates, but the strength of these step candidates is weak. The Nonparametric approach suggests an inverted U with a weak upward step and a stronger downward step for the serious cases. For the non-serious cases the Nonparametric approach suggests a nearly non-existent first step and a mild second step. Since the Nonparametric approach exhibited the best match with the observations it provides the strongest evidence against the guideline pattern in the observations for both serious and non-serious cases. The classification matrix results alone gave the impression that there was a failure of adherence to the guideline in the data. Figures 2.8 and 2.9 give a visual representation of the difference in adherence across sub-groups.



Height of confidence region corresponds to absolute size of the gradient at the candidate. Guideline steps are shown as as dotted lines. Lighter grey shading corresponds to the confidence interval of the first candidate step, darker grey shading to the second step.

Figure 2.9: Candidate steps and confidence intervals by approach and EMS contact sub-group.

2.7 Conclusion and discussion

This paper presents a new standardized methodology for assessing adherence to a clinical practice guideline. The framework employs both a classification matrix and nonparametric conditional density estimation approach. Several metrics for summarizing the classification matrix are considered, with the area under the receiver operator characteristics curve (AUC) and Cohen's κ being favoured. Observations are smoothed using state-of-the-art nonpara-

metric conditional density estimation. Candidates for discrete changes in the profile of estimates are identified using the gradients of the estimates. Candidate discrete changes are then evaluated for distinctness from each other, and for agreement with the discrete changes suggested by the guideline using bootstrapped confidence intervals.

The data in this paper represent a previously un-examined sub-population of volunteer non-physician emergency health practitioners. Adherence to clinical practice guidelines has previously been suggested to differ substantially between physicians and non-physicians, with non-physicians often adhering more strictly to practice guidelines Higuchi et al. (2012). In a volunteer setting adherence to guidelines impacts the quality of service provided and may very well affect volunteer satisfaction and individual confidence in carrying out their duties. Higuchi et al. (2012) studied non-physician and non-volunteer practitioners (nurses) in Timor-Leste and found that guidelines were well adhered to, and had the qualitative effect of increasing confidence in the appropriateness of care provided. Both quality of care delivered and retention of volunteer resources are critical in the setting investigated in this work. In this paper volunteers tended towards over-administration of a medical therapy with a relatively low-risk of adverse health consequences. This may imply that over-administration provided responders with a sense of helping or that their focus was upon patient well-being rather than strict adherence to the guideline. The method for dealing with missing observations also results in estimates which tend towards over estimation of over-administration, since some patients who genuinely needed oxygen but did not have a respiration rate recorded had a normal range respiration rate applied to their case. Further work in this area might include improving data collection processes. As well, administration of qualitative surveys of volunteer satisfaction and confidence in their treatment with respect to the guide-

line as well as behavioural experiments studying the effects of different training techniques upon adherence may be helpful in better understanding the decisions of this particular subset of practitioners. As well, further studies of the potential for crowding out behavior should be undertaken prior to implementation of any program which would enforce guideline adherence more strictly. In this setting it may be that a policy of strict guideline enforcement could shift MFR attention away from accurate scene assessment and overall patient well being, and towards guidelines in a manner which would be unhelpful for both patients, and the organization. This is because in many emergency situations over-administration of oxygen is in fact desirable and the consequences of over-administration small. In non-emergency settings oxygen administration is also undesirable but the costs of greater guideline adherence were found to be negligible.

This work contributes to efforts to improve quality of medical care by providing a framework for evaluating the adherence to clinical practice guidelines. Insights about over- and under- administration of a medical therapy are obtained and the framework enables estimates of the financial impact of improved guideline adherence to be made. In most cases, no new data collection is required to implement the framework. The framework fits into the larger system of health care system assessment detailed by Tugwell et al. (1985), and would readily integrate into the system of appropriateness assessment suggested by Brook (2009). Such integration would provide a constantly updated measure of adherence to guidelines which would be comparable across regions and levels of the health care system.

References

- Andritsos, D. A. and Tang, C. S. (2014). "Linking Process Quality and Resource Usage: An Empirical Analysis". In: *Production and Operations Management* 23.12, pp. 2163–2177.
- Arditi, C., Rege-Walther, M., Wyatt, J., Durieux, P., and Burnand, B. (2012). "Computer-generated reminders delivered on paper to healthcare professionals; effects on professional practice and health care outcomes". In: *Cochrane Database Syst Rev* 12.12.
- Askildsen, J. E., Holmås, T. H., and Kaarboe, O. (2011). "Monitoring prioritisation in the public health-care sector by use of medical guidelines. The case of Norway". In: *Health economics* 20.8, pp. 958–970.
- Barbui, C., Girlanda, F., Ay, E., Cipriani, A., Becker, T., and Koesters, M. (2014). "Implementation of treatment guidelines for specialist mental health care." In: *The Cochrane database of systematic reviews* 1.
- Brook, R. H. (2009). "Assessing the appropriateness of care - its time has come". In: *Journal of the American Medical Association* 302.9, pp. 997–998.
- Canadian Institute for Health Information (2014). *Drug Use among Seniors on Public Drug Programs in Canada, 2012*. Ottawa, ON.
- Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.
- Dimakou, S., Parkin, D., Devlin, N., and Appleby, J. (2009). "Identifying the impact of government targets on waiting times in the NHS". In: *Health care management science* 12.1, pp. 1–10.
- Fiander, M., McGowan, J., Grad, R., Pluye, P., Hannes, K., Labrecque, M., Roberts, N., Salzwedel, D. M., Welch, V., and Tugwell, P. (2015). "Interventions to increase the use of electronic health information by health

- care practitioners to improve clinical practice and patient outcomes". In: *Cochrane Database of Systematic Reviews* 3.
- Flodgren, G., Pomey, M.-P., Taber, S. A., and Eccles, M. (2011). "Effectiveness of external inspection of compliance with standards in improving health-care organisation behaviour, healthcare professional behaviour or patient outcomes". In: *Cochrane Database Syst Rev* 11.
- Flodgren, G., Conterno, L., Mayhew, A., Omar, O., Pereira, C. R., and Sheperd, S. (2013). "Interventions to improve professional adherence to guidelines for prevention of device-related infections". In: *Cochrane Database Syst Rev* 3.
- Frey, B. S. and Jegen, R. (2001). "Motivation crowding theory". In: *Journal of economic surveys* 15.5, pp. 589–611.
- Hall, P., Racine, J. S., and Li, Q. (2004). "Cross-validation and the estimation of conditional probability densities". In: *Journal of the American Statistical Association* 99, pp. 1015–1026.
- Hayfield, T. and Racine, J. S. (2008). "Nonparametric Econometrics: The np Package". In: *Journal of Statistical Software* 27.5.
- Health Council of Canada (2009). *Value for Money: Making Canadian Health Care Stronger*. Tech. rep. Toronto: Health Council, p. 52.
- Higuchi, M., Okumura, J., Aoyama, A., Suryawati, S., and Porter, J. (2012). "Application of Standard Treatment Guidelines in Rural Community Health Centres, Timor-Leste". In: *Health policy and planning* 27.5, pp. 396–404.
- Institute of Medicine (2006). *Medicare's Quality Improvement Organization Program: Maximizing Potential (Series: Pathways to Quality Health Care)*. Washington, DC: The National Academies Press. URL: <http://www.nap.edu/catalog/11604/medicares-quality-improvement-organization-program-maximizing-potential-series-pathways-to>.

- Jeni, L., Cohn, J. F., and De La Torre, F. (2013). "Facing Imbalanced Data—Recommendations for the Use of Performance Metrics". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, pp. 245–251.
- MacKinnon, J. G. (2007). *Bootstrap Hypothesis Testing*. Working Papers 1127. Queen's University, Department of Economics. URL: <https://ideas.repec.org/p/qed/wpaper/1127.html>.
- McGlynn, E., Asch, S., Adams, J., Keeseey, J., Hicks, J., DeCristofaro, A., and Kerr, E. (2003). "The Quality of Health Care Delivered to Adults in the United States". In: *New England Journal of Medicine* 348.26, pp. 2635–2645. URL: <http://dx.doi.org/10.1056/NEJMs022615>.
- Nelson, B. (2015). "Waste: Unnecessary Overuse of Medical Care Causes Both Waste and Harm". In: *The Hospitalist* 19.6:1, pp. 23–27.
- O'Brien, M., Rogers, S., Jamtvedt, G., Oxman, A., Odgaard-Jensen, J., Kristoffersen, D. T., Forsetlund, L., Bainbridge, D., Freemantle, N., Davis, D., et al. (2007). "Educational outreach visits: effects on professional practice and health care outcomes". In: *Cochrane database syst rev* 4.4.
- R Core Team (2015b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.
- Stone, C. J. (1977). "Consistent Nonparametric Regression". In: *The Annals of Statistics* 5.4, pp. 595–620.
- Straube, S. and Krell, M. M. (2014). "How to evaluate an agent's behavior to infrequent events? Reliable performance estimation insensitive to class distribution". In: *Frontiers in computational neuroscience* 8.43.

- Tape, T. G. (2015). *The Area Under an ROC Curve*. URL: <http://gim.unmc.edu/dxtests/ROC3.htm> (visited on 11/23/2015).
- The Lown Institute (2016). *About Us*. URL: <http://lowninstitute.org/home/vision-mission-history/> (visited on 01/06/2016).
- Thomas, L., Cullum, N., McColl, E., Rousseau, N., Soutter, J., and Steen, N. (1999). "Guidelines in professions allied to medicine". In: *The Cochrane Database of Systematic Reviews* 1.
- Thomas, S. (2016). "Playing by the rules? Agreement between predicted and observed binary choices." unpublished thesis chapter 1. PhD thesis. McMaster University.
- Tugwell, P., Bennett, K., Sackett, D., and Haynes, R. (1985). "The measurement iterative loop: a framework for the critical appraisal of need, benefits and costs of health interventions". In: *Journal of chronic diseases* 38.4, pp. 339–351.

2.A Observations without imputed missing values

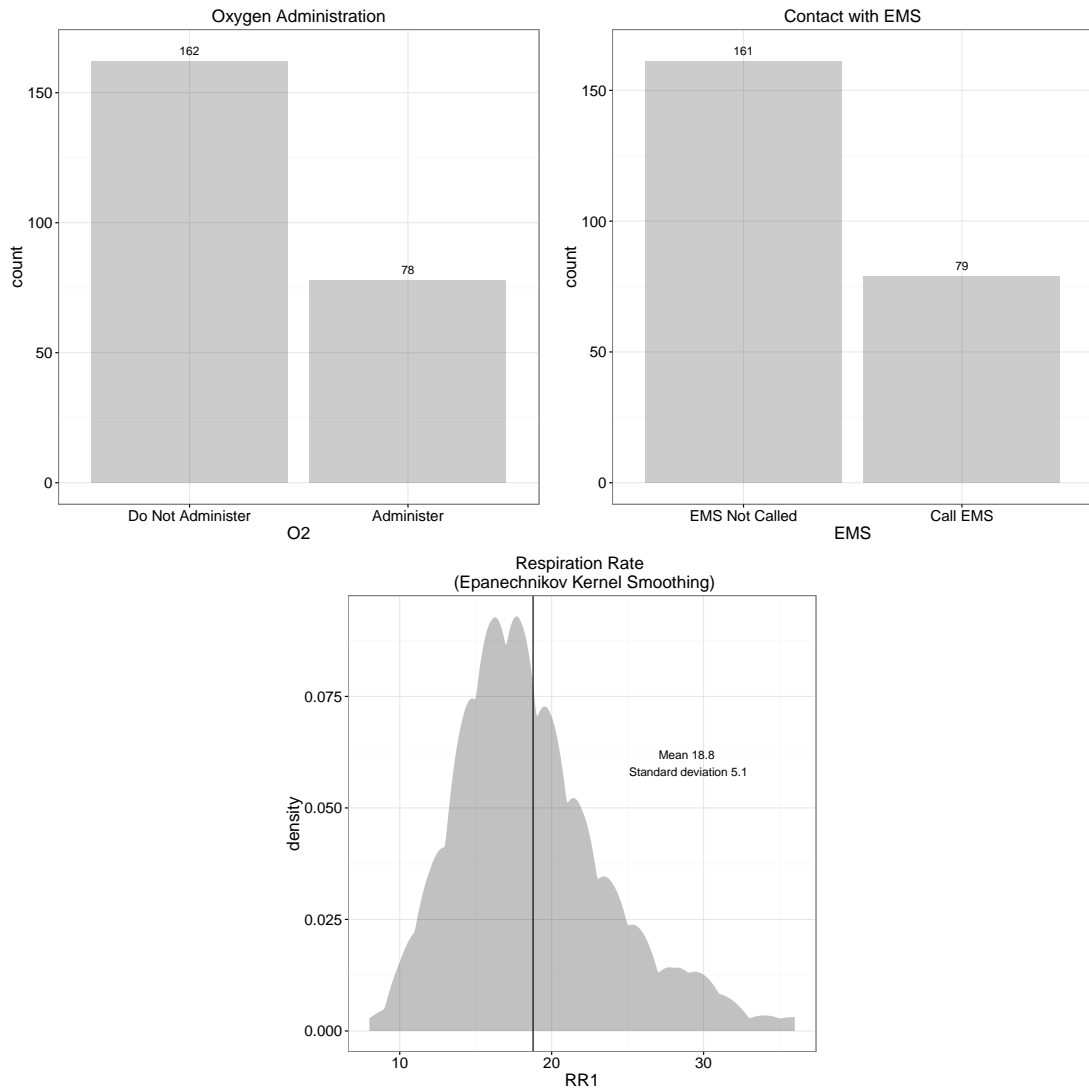


Figure 2.A.1: Illustrated data summary for data without imputed missing values.

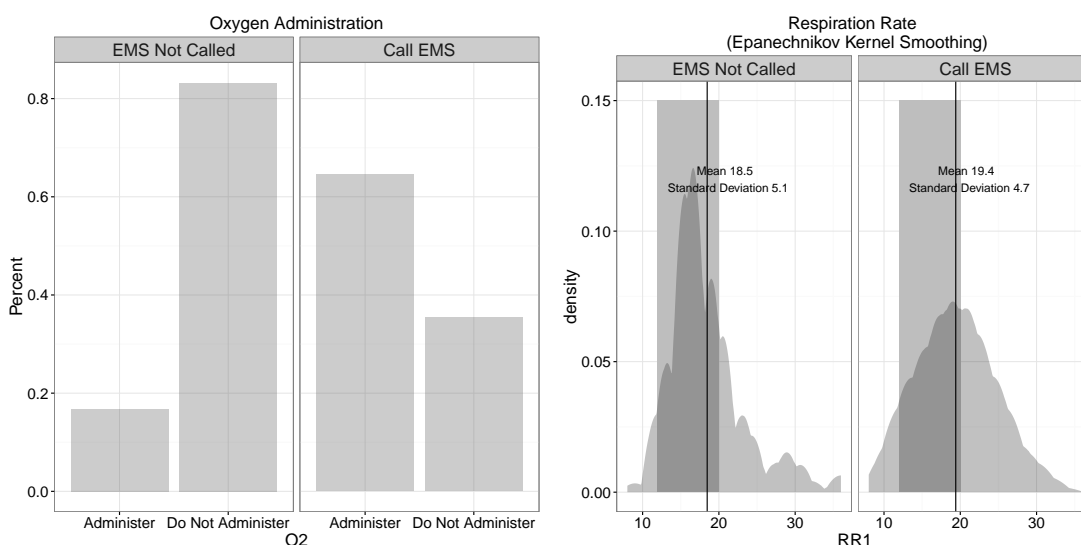


Figure 2.A.2: Illustrated data summary by EMS contact sub-group for data without imputed missing values.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	105	20	125
	Administered	29	7	36
	Total	134	27	161

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 2.A.1: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is not called for data without imputed missing values.

		Observed Oxygen Use		
		Not Administered	Administered	Total
Guideline	Not Administered	17	30	47
	Administered	11	21	32
	Total	28	51	79

Cells indicate the number of recorded cases which agree (disagree) with the respiration rate guideline for administration of supplemental oxygen.

Table 2.A.2: Classification matrix of observed oxygen administration outcomes and guideline expectations when EMS is called for data without imputed missing values.

Measure	EMS Not Called	EMS Called
TPR	25.93%	41.18%
TNR	78.36%	60.71%
FPR	21.64%	39.29%
CCR	69.57%	48.1%
bCCR	52.14%	50.95%
AUC	0.52	0.51
Cohen's κ	3.78%	1.64%

Table 2.A.3: Classification matrix summary measures by EMS contact sub-group for data without imputed missing values.

Measure	EMS Not Called	EMS Called
All Cases		
Treated Patients	5.4	10.2
Total Costs	\$ 96.92	\$ 183.08
Cost/patient	\$ 17.95	\$ 17.95
Guideline Non-Adherent Cases		
Over-Administration		
Treated Patients	4	6
Total Costs	\$ 71.8	\$ 107.7
Under-Administration		
Untreated Patients	5.8	2.2
Total Costs	\$ 104.11	\$ 39.49
Perfect Guideline Adherence		
Treated Patients	7.2	6.4
Total Cost	\$ 129.23	\$ 114.87
% Savings under Perfect Adherence	-33%	37%

Table 2.A.4: Annual cost of oxygen delivery by EMS contact sub-group for data without imputed missing values.

2.B Nonparametric Estimation

The problem at hand is to estimate the conditional density function $g(y|x) = \frac{f(x,y)}{\mu(x)}$.

$$\hat{g}(y|x) = \frac{\hat{f}(x,y)}{\hat{\mu}(x)}, \quad (2.B.1)$$

The approach used in this paper is that described by Li and Racine (2007) where the numerator and denominator of the conditional probability function are described by:

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n K_{\gamma}(x, X_i) k_{\lambda_0}(y, Y_i) \quad (2.B.2)$$

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n K_{\gamma}(x, X_i), \quad (2.B.3)$$

with $K_{\gamma}(x, X_i)$ and $k_{\lambda_0}(y, Y_i)$ representing kernel density functions.

For the purposes of this study the kernel suggested by Li and Racine (2003) will be used for the estimation of the unordered discrete variable $O2$:

$$\begin{aligned} k_{\lambda_0}(y, Y_i) &= l(Y_{is}, y_s, \lambda_s) \\ &= \begin{cases} 1 - \lambda_s & \text{if } Y_{is} = y_s \\ \frac{\lambda_s}{c_s - 1} & \text{if } Y_{is} \neq y_s \end{cases}, \end{aligned} \quad (2.B.4)$$

where y_s can take on c_s ordered values $0, 1, c_s - 1$. If $\lambda_s = 0$ then $l(Y_{is}, y_s, \lambda_s) = 1$ is an indicator function, and if $\lambda_s = \frac{c_s - 1}{c_s}$, then $l(Y_{is}, y_s, \frac{c_s - 1}{c_s}) = \frac{1}{c_s}$, a constant. Thus the range for the smoothing parameter associated with *participate* is $[0, \frac{2-1}{2} = 0.5]$.

$K_{\gamma}(x, X_i)$ is a product kernel, in this case composed of the kernel proposed by Li and Racine (2003) for the unordered levels of *EMS* and the kernel proposed by Epanechnikov (1969) for the continuous variable *RR1*.

$$K_{\gamma}(x, X_i) = W_h(x^c, X_i^c) L(x^d, X_i^d, \lambda), \quad (2.B.5)$$

where $\gamma = (h, \lambda)$ is a vector of continuous and discrete bandwidths in this case for *RR1* and *EMS*. The superscript c denotes the continuous variable *RR1* and d the discrete variable *EMS*. The Epanechnikov kernel used here is fur-

ther defined by:

$$W(u) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2) & \text{if } u^2 < 5 \\ 0 & \text{otherwise} \end{cases} \quad (2.B.6)$$

where $u = \frac{x^c - X_i^c}{h}$
and $h > 0$,

and the Li and Racine kernel by:

$$L(x_i^d, x^d, \lambda) = \begin{cases} 1 & \text{if } |x_i^d - x| = 0, \\ \lambda^{|x_i^d - x|} & \text{if } |x_i^d - x| \geq 1 \end{cases}, \quad (2.B.7)$$

where λ must lie between 0 and 1.

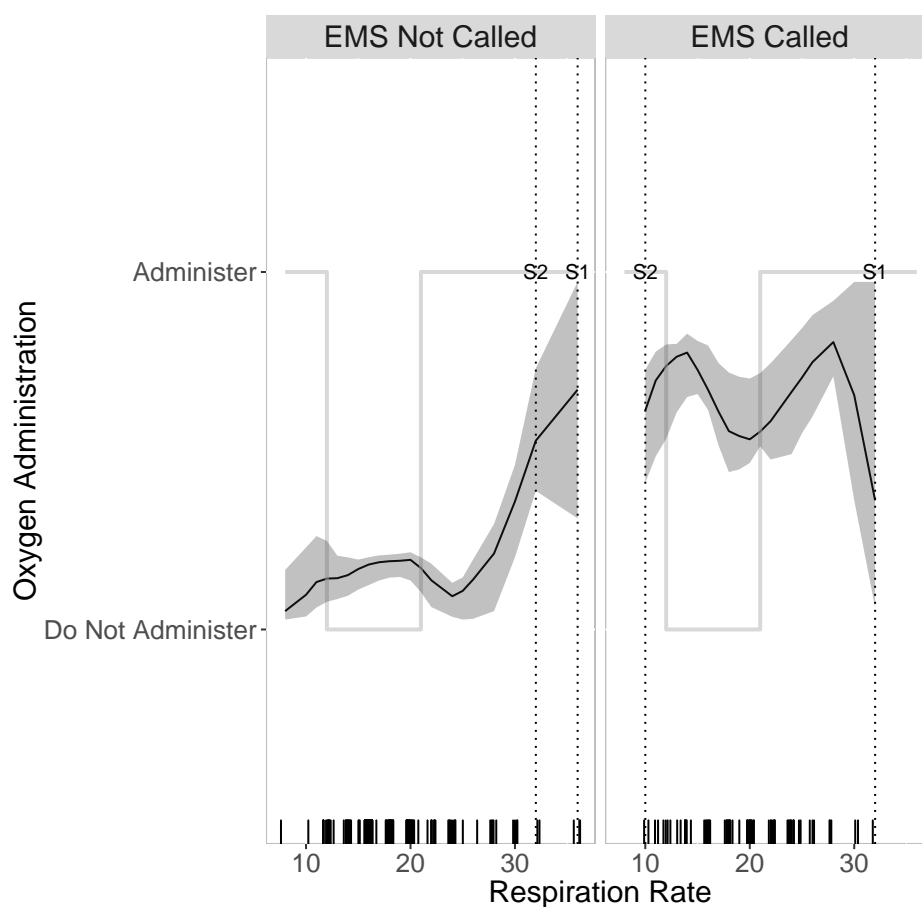
In order to estimate the nonparametric model, optimal bandwidths were determined using a least squares cross validation routine. This approach has the advantage that if a regressor is irrelevant¹³ it will be smoothed entirely out of the regression asymptotically. Smoothing out in this case is demonstrated by a large bandwidth. Table 2.B.1 provides the results. In both data sets all variables are relevant as these are well below their upper bounds.

Variable	Original	Imputed	Maximum.Range
O2	0.0285	0.0042	1
RR1	1.8510	3.3319	28
EMS	0.0000	0.0000	1

Table 2.B.1: Bandwidths generated using least squares cross validation for data with and without imputed missing values.

2.C Nonparametric results with missing values

¹³Meaning that the regressor does not substantially affect the outcome.



Guideline is the grey line. Candidate step 1 (S1) and step 2 (S2) are dotted lines. Bootstrapped 90 percent confidence interval is the shaded area. Bars along the x-axis indicate frequency of observations.

Figure 2.C.1: Estimated probability of oxygen administration by respiration rate and EMS contact sub-group using the Nonparametric approach for data without imputed missing values.

EMS	Step	Lower ci	Upper ci	Guideline	Gradient	
EMS Not Called	Step 1	36	16	36	12	-0.0900
EMS Not Called	Step 2	32	9	32	20	0.2066
EMS Called	Step 1	32	16	32	12	-0.3372
EMS Called	Step 2	10	10	28	20	0.1418

Table 2.C.1: Candidate steps of the nonparametric strategy for data without imputed missing values and bootstrapped 90 percent confidence intervals.

2.D Data sharing agreement

6/9/2015

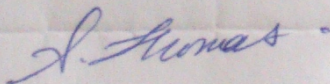
Stephanie Thomas
175 Hunter Street East #205
Hamilton ON L8N 4E7

Brent Schriener
Unit Chief
St John Ambulance Unit D0007
65 Nebo Road
Hamilton ON L8W 2C9

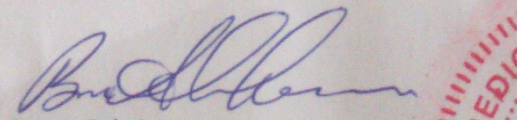
To whom it may concern,

This letter confirms the transfer of anonymized PCR data from St John Ambulance Unit D0007 to Stephanie Thomas for research purposes and authorizes the submission of the results of said research for publication. Stephanie Thomas agrees to maintain the security and confidentiality of the data accordingly and to return or destroy the data at the request of the current or future chief of St John Ambulance Unit D0007.

Sincerely,



Stephanie Thomas
BA, MA, Candidate: PhD Economics



Brent Schriener
Unit Chief St John Ambulance Unit D0007



Chapter 3

Cap and Trade versus Intensity

Targets: Emissions trading markets with stochastic demand

3.1 Introduction

Air pollutants have been linked to negative externalities ranging from reduced visibility to illness and death. Recently revised figures from the OCED estimate the global health costs of illness and death associated with air pollution to be 1.7 trillion USD in 2010 (OECD, 2014). The Intergovernmental Panel on Climate Change 2014 Synthesis Report (2014, p.8) states that

“Continued emission of greenhouse gases will cause further warming and long-lasting changes in all components of the climate system, increasing the likelihood of severe, pervasive and irreversible impacts for people and ecosystems. Limiting climate change would require substantial and sustained reductions in greenhouse gas emissions which, together with adaptation, can limit climate change risks.”

In response, many jurisdictions have begun to develop mechanisms for controlling air pollutants.

The assignment of property rights to air pollutants and the establishment of trading markets has taken place in many areas including, California, Quebec, Alberta, the European Union, Japan, Australia, and New Zealand. Efforts to develop appropriate markets in Ontario, China, India and Mexico are ongoing. While there are a multitude of ways to control air pollutants, including taxes, direct regulations and markets, the focus of this paper is on two commonly used market mechanisms: Cap and Trade and Intensity Targets (also referred to as Baseline and Credit trading, Emission Reduction Credits and Tradable Performance Standards). Under a Cap and Trade system organizers set a desired total cap on air pollutants that is substantially lower than the current emission level and assign firms emission permits equal to the cap. Firms can then redeem permits for emissions created during a redemption period for emissions they have been unable to eliminate, and sell unused permits in a permit market, if they have reduced their emissions below their allowable caps. Typically the sales will be to firms whose emission abatement costs are greater than the costs incurred by the firms selling permits. In this way emissions are reduced to the regulated level at least cost and provide a gain to the people who had been suffering the effects of the excessive pollution. The EU-Emission Trading Scheme, for example, is a Cap and Trade system. India's PAT initiative, and Alberta's Specified Gas Emitters Regulation on the other hand, are intensity based systems. Under Intensity Targets a target intensity level of air pollutants (emissions) per unit of output is established and firms with clean technology and emission rates below the target generate permits which can then be traded with firms which retain technologies which fail to meet the target and produce greater emissions per unit of output than the target. Firms with production technologies consistent with emission rates per unit of output exactly matching the prescribed intensity target are viewed as being compliant regardless of their output production

level.

While economic theory clearly supports Cap and Trade as an efficient and lower cost market regulation mechanism than Intensity Targets, theory does not always hold in practice. Both systems are observed to be employed in efforts to contain emissions. While Cap and Trade systems appear to offer predictable results in terms of total air pollutants produced, Intensity Targets offer policy makers the advantage of avoiding the assignment of permits to firms and are often supported by industry due to the lack of a binding cap on emissions. Under a Cap and Trade system emission permits are typically initially assigned by either granting permits in proportion to firm emissions prior to the regulation, in a process referred to as 'grandfathering', or by auctioning the available permits. Firms which have already taken steps to reduce their emissions will not benefit in the same way that firms who make these reductions after the allocation of permits will benefit when permits are allocated using grandfathering. Similarly, when permits are allocated using an auction firms with fewer resources to purchase permits may be crowded out of the market by high permit prices driven by demand from high emitting firms, or even relatively clean firms if they are large firms. In addition, it may be difficult to accurately identify firm emissions when estimating firm size for permit allocation. An additional concern which led to the favouring of an intensity based system over a Cap and Trade system in India was expressed by the Indian government which held the view that climate change was caused by developed countries. In this case the focus on improving production efficiency circumvented the need to establish consensus on a cap (Upadhyaya, 2010) and the country began to pilot the 'Perform Achieve and Trade (PAT)' emissions system in 2012.

While both Cap and Trade and Intensity Target systems are observed in practice, applied research about the performance of these systems in the field is

scarce. Newly established systems are too young to offer the data necessary to effectively assess the issue empirically. As well, comparisons across jurisdictions are limited due to confounding factors such as differences in implementation, culture and industrial conditions. This makes the experimental economics approach to assessment of the performance of Cap and Trade and Intensity Target market mechanisms particularly useful. Cason and Plott (1996) demonstrated the value of experimental methods in the design of the United States Environmental Protection Agency's emission trading markets, leading to improvements in the design of sulphur dioxide markets. Here we compare the performance of Cap and Trade to Intensity Targets in the control of industrial pollutants to explore the relative merits and behavioural responses under each system in a controlled laboratory setting.

In this paper Cap and Trade is compared directly to Intensity Targets in a controlled laboratory setting. This work builds upon Buckley, Mestelman, and Muller (2014), in which firms can adjust both their abatement technology and output in response to market conditions. In our new environment, however, demand for firm output and the prices of emission permits are uncertain. Unlike Buckley et al. (2014), instead of equating the experimental parameters so that average production intensity is equivalent in both markets (resulting in higher aggregate emission levels under Intensity Targets compared to Cap and Trade), we create a setting in which both systems are expected to result in the same level of emissions. While firms still choose their production technology and level of output, in this work we add uncertainty in the demand for output to better incorporate the market conditions faced by firms operating in the field. This allows us to study these market instruments in a more realistic stochastic environment which will potentially cause inefficiencies to Cap and Trade compared to Intensity Targets, due to the fixed nature of the cap as a quantity based instrument. In addition, we

study the effects of allowing banking of permits under each of the Cap and Trade and Intensity Target systems. While banking of emission permits is necessary for firms facing stochastic demand under Cap and Trade regulation, banking is not necessary to achieve a market-clearing equilibrium under an Intensity Target regulation due to the fact that permit supply is endogenous and fluctuates with the level of economic activity because the regulated target is linked to emission intensity per unit of output. However, in a similar environment Buckley et al. (2014) found that participants tended to keep more permits than necessary over the lifespan of the experiment. We also incorporate uncertainty about continuation of the policy by having a random end-date. All of these modifications serve to improve the external validity of the comparison of Cap and Trade and Intensity Target mechanisms to emissions control.

Our objectives are to study the effects of uncertainty on firm choices. Specifically, we study the effect of uncertainty in demand for output expressed as uncertainty in the returns on each unit of production. We also study the effect that this demand uncertainty has on the market for permits due to the fact that output demand shocks will implicitly effect the value of firm abatement decisions. Firms can respond to uncertainty by adjusting on two margins in this environment. By changing to a cleaner or dirtier abatement technology firms adjust on the intensity margin. On the production margin firms can respond by changing the level of output produced. In response to uncertainty we would like to know if firms favour one margin of adjustment over another. We implement the two emission systems in the laboratory environment so that the aggregate emissions in each of the two systems is expected to be identical. This is done so that we may address the concerns of policy makers aiming to choose between these two systems in order to meet mandated emissions targets and to explicitly test if this equivalence is borne out

in practice. We test whether the relatively high permit prices and emission permit inventories found under baseline and credit trading in previous single margin experiments are borne out in an environment more similar to the field and investigate whether or not agents optimize using similar techniques in both systems. We explore whether specific firm types as defined by the optimal level of emissions intensity as 'clean' or 'dirty' firms are responsible for expansion of emissions, as was found in Buckley et al. (2014), or whether expansion is proportional. As well, while banking is not predicted to be necessary under Intensity Targets, we assess whether agents bank permits in response to uncertainty.

3.2 Literature

Many researchers have investigated Cap and Trade and Intensity Targets as market based methods for controlling emissions. The theoretical static comparison of Cap and Trade with Intensity Targets is presented in the work of Dewees (2001), Fischer (2001) and Fischer (2003). All three find that Intensity Targets are inefficient relative to Cap and Trade. However, relative performance can be influenced in theory by market structure, as found by Boom and Dijkstra (2009), by uncertainty in emissions relative to output and by marginal abatement costs, as presented in Quirion (2005), and by growth rates as described by Tian and Whalley (2009). As well, Jotzo and Pezzey (2007) find that Intensity Targets result in lower variance in abatement costs when there is uncertainty in future permit allocations.¹ Marschinski and Lecocq (2006) expand the range of uncertainty, including uncertainty in future GDP, future business as usual emissions and future abatement costs confirming the

¹Although, Tian and Whalley (2009) and Jotzo and Pezzey (2007) specify emission intensity as a country's aggregate emissions relative to GDP, which is different from how we describe emission intensity in this paper.

finding of Jotzo and Pezzey (2004) that variance in abatement costs is lower under Intensity Targets than Cap and Trade. From a dynamic perspective, Fischer and Springborn (2011) propose a macroeconomic real business cycle model which compares intensity targets to Cap and Trade and suggest that labor, capital and output are higher under Intensity Targets than either taxes or a cap, and no more volatile than without a policy. Stranlund, Murphy, and Spraggon (2014), on the other hand, model permit prices under Cap and Trade, incorporating banking and price controls into an industry-level stochastic dynamic model with uncertain future production. Combined with an experiment these authors find that permit prices are less volatile with banking than without, with price controls than without, and lowest with a combination of both price controls and permit banking. Both of these models suggest that there may be a larger role for dynamic investigation of permit markets.

The use of economic experiments to investigate emission permit markets is now well established. However, no experiments exist to our knowledge, which directly compare Cap and Trade to firm-level intensity choices while incorporating uncertainty in demand for output. Direct experimental comparisons of Cap and Trade to Intensity Targets are carried out in a series of papers by Buckley, Mestelman, and Muller (2006), Buckley, Mestelman, and Muller (2008), and Buckley et al. (2014), to which this represents an extension. These papers build a tractable model of firm behaviour and implement a series of experiments which progressively increase the complexity of decisions faced by lab participants. Buckley et al. (2006) first allow both intensity choice and output capacity choice for robot traders, and allow human participants a choice of technologies under fixed output capacities. Buckley et al. (2008) then allow participants to make output capacity choices under fixed intensity choices. Finally, Buckley et al. (2014) extend both intensity and

output choices to student participants. Experiments which incorporate uncertainty have focused only on Cap and Trade programs. Among these are Godby, Mestelman, Muller, and Welland (1997), Godby, Mestelman, Muller, and Welland (1998), and Cason and Gangadharan (2006) who incorporate uncertainty in the amount of emissions produced by a firm; and Ben-David et al. (2000), who incorporate uncertainty in the timing of a reduction of the aggregate cap, and compare this to uncertainty in the amount of reduction in the aggregate cap. These authors also incorporate a role for risk aversion to influence firm decisions to adopt more efficient production technologies.² While all the experiments cited here involve subjects choosing the number of permits to buy and sell in a permit market, in all but Buckley et al. (2014), firm choice is limited to choosing abatement levels by choosing an absolute emission level, as in Cason and Gangadharan (2006) or to a production technology, as in Ben-David et al. (1999) and Ben-David et al. (2000). Only in the environment presented in Buckley et al. (2014) are live participants offered the choice of making decisions on the emission intensity margin and the output production margin.³ While Buckley et al. (2014) allow only single unit adjustment of firm output from period to period, the experiment described in this paper affords participants, acting as firms, the choice of both emission intensity(i.e. production technology) and output production levels across a full range of values. Participants face uncertainty in demand in output markets, and must make decisions on both emission intensity and the amount of output to produce.

The impact of permit banking has been explored within the context of Cap

²Other authors specify aggregate caps which are set relative to GDP (Sue Wing, Ellerman, and Song, 2006), in which case uncertainty about country level growth rates becomes an important, as in Tian and Whalley (2009). Our specification of intensity is at the firm level so these emissions/GDP type approaches are beyond the scope of this paper.

³Buckley et al. (2006) employed robot traders who could choose both production emission intensity technology and output capacity.

and Trade emission permit trading. Muller and Mestelman (1994) and Mestelman, Moir, and Muller (1999) consider the impact of banking and trade of emission permits and shares (a stream of emission permits) upon emission permit prices and efficiency under two different market institutions. These authors compare results to similar experiments run at other sites, finding relatively stable permit prices and higher efficiency than at other experiment locations, but that banking does not lead to gains above a trading only equilibrium. Godby et al. (1997), and Godby et al. (1998) incorporate emissions uncertainty into their experimental designs finding that uncertainty contributes to price instability which is resolved with banking and borrowing provisions. Cason and Gangadharan (2006) and Stranlund, Murphy, and Spraggon (2011) study the interaction of permit banking with compliance and uncertainty in emissions, and uncertainty in the level of the aggregate cap, respectively. Ben-David et al. (1999) and Ben-David et al. (2000) both allow permit banking as participants choose irreversible emission intensity levels, but this is not the core focus of these studies.⁴ Stranlund et al. (2014) study the impacts of banking upon permit prices incorporating uncertainty about future production levels, as well as in the termination of the program. The design of the experiment reported in this paper is a combination of the emission intensity decision modeled by Buckley et al. (2014) and the output decision modeled by Stranlund et al. (2014). These elements meet the minimum requirements to test for differences between Intensity Target and Cap and Trade systems while maintaining out study's comparability to the existing literature.

⁴Murphy and Stranlund (2006) and Murphy and Stranlund (2007) also explore compliance in experimental emission permit markets, but exclude banking.

3.3 Experimental design

A fully specified environment with a double auction emission permit market, an explicit emission technology choice, the ability of firms to bank unused permits for future periods, and uncertainty in output demand is required to test our theoretical predictions concerning the alternative emission trading plans based on an established emission cap or intensity targets in a realistic long-run setting. In order to focus on market features important to our theoretical predictions, the experimental setting necessarily abstracts from many additional market characteristics that would exist in a naturally occurring setting. Failure to abstract makes it difficult to focus on how the incentives directly associated with the specific plans to regulate emissions will affect emissions. Thus, we impose full compliance, abstracting from issues of penalties and monitoring. Compliance is enforced by restricting the output decision based on the subject's current holding of emission permits. Firms are not able to sell output if they do not have the required amount of permits to provide to the regulator. Bankruptcy is prevented by requiring firms to hold sufficient earnings to cover the cost of permit purchases and by not allowing production decisions which result in non-positive total earnings.

Subjects are told that they represent firms that require inputs to create output. Inputs can be bought from and sold to other participants in the experiment. In the Cap and Trade treatments subjects are given an endowment of inputs at the start of each period. We present the experiment in neutral terms because we want to explore the economic incentives of the problem independent of the social context, which may result in decisions based on personal preferences for environmental activity rather than responses to costs. Feedback from pilot sessions indicated that our terminology was sufficiently neutral that subjects did not infer emissions permits to be the underlying motiva-

tion for the structure of the experiment.

We employ a design using two types of firms, each possessing a different marginal abatement cost (MAC) schedule. Type A firms are 'clean' and have relatively flat MAC schedules while the Type B firms are 'dirty' and have relatively steep MAC schedules. Thus, it is cheaper for 'clean' firms to abate pollution than it is for 'dirty' firms. Here we use the terminology of 'clean' versus 'dirty' to denote that it is cheaper for clean firms to abate emissions but the ultimate cleanliness and emissions levels of these firms is a decision left up to each participant. There are four subjects of each type in each session. Section 3.4 outlines the underlying theory and Section 3.5 details the implementation of the experiment.

3.4 Theory

The theoretical model is based on the dynamic stochastic emission permit trading model of a Cap and Trade environment presented in Stranlund et al. (2014). Demand for output is stochastic and a fixed marginal benefit schedule is presented to participants, however no emissions intensity decision is required within the framework of Stranlund et al. (2014). We extend this environment, borrowing from Buckley et al. (2014) to incorporate an emissions intensity decision such that there is a clear relationship between the marginal benefit schedules and this emission intensity. The environment is consistent with an industry with a fixed number of firms producing output and emissions. Each firm faces a market for output which is distinct from all other firms, and so operates as a monopolist on the output margin. Simultaneously, each firm participates in a common competitive emissions trading market. The monopoly nature of the output market means that it is possible to deduce output demand schedules from the marginal benefit schedules pre-

sented to participants in the experiment. These output demand schedules are sufficient to produce the marginal benefit schedules presented to participants and are consistent with decreasing demand for output and constant marginal costs of production. This also allows one to form a complete assessment of the total surplus generated in each treatment of the experiment.⁵

We first present the stochastic dynamic model which reduces to a static model when banking is not permitted. As well, under Intensity Targets clearing of the permit market in each period implies that banking is not necessary for optimization, and so the dynamic formulation applies only to the case of Cap and Trade with Banking. Next, the static model is presented.

3.4.1 A dynamic stochastic model of emission permit trading with banking

The model of Stranlund et al. (2014) combines the work of Schennach (2000), Fell and Morgenstern (2010) and Fell, Burtraw, Morgenstern, and Palmer (2012) to study the effects of banking and price controls in a Cap and Trade emissions market with fixed firm intensities. It forms the basis of our dynamic stochastic model to study the effect of banking upon equilibrium emission permit prices in a Cap and Trade system of emission permit regulation. We extend this model to incorporate multiple levels of emissions intensity and compare it with Intensity Target regulation.

Suppose there is an industry composed of N firms such that $i \in [1, \dots, N]$, operating in time periods $t = 1, \dots, t, \dots$, who experience an industry wide stochastic shock, u_t , affecting demand for output, $u_t \in [low, high]$. A competitive market for emission permits enables an industry level model of dynamic

⁵It is also possible to deduce a demand schedule consistent with the marginal benefit schedule in which output prices are constant and marginal costs are increasing, however this leads to an indeterminate consumer surplus and is difficult to conceptualize in a real world industrial setting.

permit trading. Let $\pi_t(\Omega_t, u_t)$ be the maximum industry profits given Ω_t total emissions, with $\pi_{t,\Omega_t}(\Omega_t, u_t) > 0$ and $\pi_{t,\Omega_t\Omega_t}(\Omega_t, u_t) < 0$. The output demand state u_t is unknown in periods before t . The likelihood and magnitude of each possible demand state is public information within the experiment, with the state itself revealed to participants at the beginning of t .

Regulation allows firms to bank permits for future use or sale, but does not allow borrowing from future allocations. Under Cap and Trade regulation this means that the industry's stock of banked permits at the beginning of $t + 1$ is S_{t+1} , and the evolution of the aggregate bank of permits is

$$S_{t+1} = S_t + A_t - \Omega_t \geq 0, \quad (3.1)$$

where A_t is the industry allocated emissions permits in period t . Under Intensity Targets the regulator sets an industry wide target emissions intensity rate r^T (expressed in terms of emissions per unit of output) which is insensitive to output demand. Each firm responds to this target by choosing an optimal emission intensity r_{i,u_t} in time t , prior to the revelation of the shock u_t . After the revelation of the shock each firm also chooses an output q_{i,u_t} . The evolution of the aggregate bank of permits under Intensity Targets is therefore

$$S_{t+1} = S_t + \sum_{i=1}^N (r_{i,u_t} - r^T) q_{i,u_t} \quad (3.2)$$

In equilibrium, competitive emission permit markets clear, thus

$$\sum_{i=1}^N r_{i,u_t}^{IT} q_{i,u_t}^{IT} = \sum_{i=1}^N r^T q_{i,u_t}^{IT} \quad (3.3)$$

which implies that banking is not necessary for optimization under Intensity Targets. This is because, unlike under Cap and Trade, the supply of permits

is endogenous under Intensity Targets. While Cap and Trade regulation constrains absolute emission levels, Intensity Target regulation only constrains emission level relative to output. Thus, only Cap and Trade requires firms to bank permits in order to expand output when it is favourable to do so. When demand is high (low) clean firms expand (contract) output to create (reduce) the supply of permits required by dirty firms who also expand (contract). This implies that the dynamic model refers only to the Cap and Trade setting in which firms must bank permits to maximize profits when demand is uncertain.

As stated above, borrowing from the model presented in Stranlund et al. (2014), we also assume that in every period there is an exogenous probability, γ , that the session will continue into the next period. $\gamma = 1$ for the first 10 periods and then $\gamma < 1$ for the remaining periods. The time paths of expected emissions, banking, and permit prices can be determined from the stochastic dynamic programming problem of choosing $(\Omega_0, \Omega_1, \dots)$ such that $\Omega_t = \sum_{i=1}^N r_{i,u_t} q_{i,u_t}$ to maximize the expected present value of industry profits, such that industry profits in each period are defined by $\pi(\Omega_t, u_t)$.

$$E_0 \left[\sum_{t=0}^{\infty} \gamma_t (1 + \mu)^{-t} (\pi(\Omega_t, u_t)) \right] \quad (3.4)$$

$$\text{such that: } S_{t+1} = S_t + A_t - \Omega_t,$$

$$\text{and } S_0 = 0$$

Market equilibrium in a period requires that $\pi_{\Omega_t}(\Omega_t, u_t) = p_t$, where p_t is the mean price of an emission permit in period t and μ is the constant discount rate. $E_0[\cdot]$ denotes the expected value at the beginning of the program. Thus given a realization of u_t , the competitive price and aggregate emissions are inversely related. Given a non-empty permit bank $S_{t+1} > 0$, there is a positive relationship between current permit price and the size of the bank at the

end of the period. Higher prices imply lower aggregate emissions and more permit banking.

Solving the stochastic dynamic programming problem, as in Stranlund et al. (2014), results in

$$p_t^{CT} = \lambda_t + \sum_{s=1}^{n-1} \left(\prod_{s=1}^s \gamma_{t+s} (1 + \mu)^{-s} E_t[\lambda_{t+s}] \right) + \prod_{s=1}^n \gamma_{t+s} (1 + \mu)^{-n} E_t[p_{t+n}^{CT}], \quad (3.5)$$

relating permit price in time t with expected price from time n in the future, where $\lambda_\tau \geq 0$, $\forall \tau = t, \dots, t+n$ are the Lagrange multipliers attached to the no-borrowing constraints in each period.

If $\gamma = 1$ in the interval $t, t+n$, so there is no chance that the session will end, and the permit bank is zero then $E_t[\lambda_{t+s}] = 0$ for each $s = 0, \dots, n$ and $p_t^{CT} = (1 + \mu)^{-n} E_t[p_{t+n}^{CT}]$, which implies that the expected price of permits increases at the rate of discount.

On the other hand, if participants in the experiment do not discount over the time frame of the experiment prices might fall as the experiment progresses because

$$p_t^{CT} = \lambda_t + \sum_{s=1}^{n-1} E_t[\lambda_{t+s}] + E_t[p_{t+n}^{CT}] \quad (3.6)$$

which implies that $p_t^{CT} > E_t[p_{t+n}^{CT}]$.

As mentioned by Stranlund et al. (2014), if there is an extended time in which the session will continue with certainty, participants are expected to bank permits in early periods and draw down this bank in later periods. When the session is no longer certain the incentive to bank is reduced by $\gamma < 1$. This reduced incentive to bank reduces prices, increases emissions and reduces permit banking.

3.4.2 A static model of emission permit trading

There are three cases in which the problem of optimization reduces to one of static optimization which implies that expected permit prices p_t are constant over all periods: Cap and Trade without Banking, Intensity Targets with Banking and Intensity Targets without Banking. In all cases all firms face industry wide stochastic shocks to output demand $u \in [low, high]$. In each period, prior to revelation of the demand state, each firm chooses an emission intensity $r_{i,u}$, which is similar to firms calibrating their production facilities prior to producing their product. These kinds of technological choices tend to be medium to long-run decisions that must be made before exact demand conditions are known. Once the emission intensity has been chosen, the output demand state is revealed and firms choose $q_{i,u}$ for the period. A firm's total emissions in a particular state are $e_{i,u}$, such that $e_{i,u} = r_{i,u}q_{i,u}$; thus emissions intensity is simply the rate of emissions per unit of output. Industry output is $Q_u = \sum_{i=1}^N q_{i,u}$. Aggregate emissions are $\Omega_u = \sum_{i=1}^N r_{i,u}q_{i,u}$. Environmental damages are the same as in Buckley et al. (2014) with $D = D(\Omega_u)$, $D'(\Omega_u) > 0$ and $D''(\Omega_u) \geq 0$. Output demand is modeled as each firm operating as a monopolist. Aggregate revenues are thus $\sum_{i=1}^N P_{i,u}(q_{i,u})q_{i,u}$ such that each firm may have a different price, and choose a different output, in each output demand state.

The private cost of production is a linear homogeneous function of output and intensity: $C_i = C_i(q_{i,u}, e_{i,u}) = q_{i,u}C_i(1, r_{i,u})$. Unit cost $C_i(1, r_{i,u})$ can be separated into unit capacity cost $c_i(r_{i,u})$, which is a positive and declining function of the emission intensity with $c_i(r_{i,u}) > 0$ and $c'_i(r_{i,u}) \leq 0$, and unit variable cost w_i which is a constant function of output. Total cost is thus $C_i = c_i(r_{i,u})q_{i,u} + w_iq_{i,u}$. Marginal cost of output is equal to $c_i(r_{i,u}) + w_i$ and the marginal cost of abating pollution is $MAC = -\frac{\partial c_i}{\partial e_{i,u}} = -c'_i(r_{i,u})$.

The social profit, \mathcal{S} , is composed of firm revenues net of production costs and damages. The emission regulator's problem is to permit firms to maximize profits while taking into account the environmental damages of their activities by choosing $r_{i,u}, q_{i,u}$ for each firm in each output demand state. The outcome of Equation 3.7 is defined as the social profit maximizing (SPM) outcome because it incorporates social damages.

$$\max_{\{r_{i,u}, q_{i,u}\}} \mathcal{S} = E_u \left[\sum_{i=1}^N P(q_{i,u})q_{i,u} - \sum_{i=1}^N (c_i(r_{i,u})q_{i,u} - w_i q_{i,u}) - D\left(\sum_{i=1}^N r_{i,u}q_{i,u}\right) \right] \quad (3.7)$$

where E is the expected value over all possible demand states.

The first order conditions for an interior maximum are:

$$-E_u [c'_i(r_{i,u}^*)] = E_u \left[D'\left(\sum_{i=1}^N r_{i,u}^* q_{i,u}^*\right) \right] \quad \forall i \in N, \forall u \quad (3.8)$$

and

$$E_u [P'_{i,u}(q_{i,u}^*)q_{i,u}^* + P_{i,u}(q_{i,u}^*)] = E_u \left[c_i(r_{i,u}^*) + w_i + r_{i,u}^* D'\left(\sum_{i=1}^N r_{i,u}^* q_{i,u}^*\right) \right] \quad \forall i \in N, \forall u \quad (3.9)$$

with optimal values $q_{i,u}^*$ and $r_{i,u}^*$ greater than or equal to zero and

$$E_u \left[\sum_{i=1}^N r_{i,u}^* q_{i,u}^* \right] = E_u [\Omega_u^*],$$

the optimal level of expected aggregate emissions. Optimization is carried out on two margins. Equation 3.8 ensures efficient abatement because expected marginal costs of each firm are equated to the expected marginal damages of

the industry. Marginal abatement cost is equated across firms and equal to marginal damage. Equation 3.9 ensures efficient output which is social profit maximizing, because the expected return per unit of output is equated to the expected costs per unit of output for each firm.

The regulator's optimum can be supported as a competitive equilibrium under Cap and Trade regulation. The emissions regulator distributes A_i allowances to each firm, regardless of the demand state, u , such that the sum of allowances equals the optimal level of emissions, $\sum_{i=1}^N A_i = E_u[\Omega_u] = \Omega^*$. Letting $p_{CT,u}$ denote the price of allowances under Cap and Trade when banking is not permitted, firm i 's profit maximization problem is

$$\max_{\{r_{i,u}, q_{i,u}\}} E_u \left[\pi_{i,u}^{CT} \right] = E_u \left[P_{i,u}(q_{i,u})q_{i,u} - c_i(r_{i,u})q_{i,u} - w_i q_{i,u} - p_{CT}(r_{i,u}q_{i,u} - A_i) \right]. \quad (3.10)$$

The two first order conditions for an interior maximum are

$$-E_u[c'_i(r_{i,u}^{CT})] = E_u[p_{CT,u}] \quad (3.11)$$

if $q_{i,u}$ is greater than zero, and

$$E_u[P'_{i,u}(q_{i,u}^{CT})q_{i,u}^{CT} + P_{i,u}(q_{i,u}^{CT})] = E_u[c_i(r_{i,u}^{CT}) + w_i + r_{i,u}^{CT} p_{CT,u}] \quad (3.12)$$

Equation 3.11 ensures expected cost minimizing abatement and defines each $r_{i,u}$. Equation 3.12 requires that each firm earn zero expected marginal profit, and identifies each $q_{i,u}$, and therefore Q_u^{CT} . The system (3.11) and (3.12) can be obtained from the SPM (3.8) and (3.9) if $p_{CT,u}$ replaces $D'(\sum_{i=1}^N r_{i,u}^* q_{i,u}^*)$. Note that since the regulator chooses A_i based on $\Omega^* = E_u[\Omega_u]$ this means that $p_{CT,u} = p_{CT}$ for all u . Thus, the solution to the SPM problem can be sustained as a Cap and Trade competitive equilibrium and vice versa.

Under an Intensity Target plan, the regulator sets an industry-wide emission intensity target, r^T , which is insensitive to changes in output demand. This emission intensity target characterizes a relative emission target mechanism. Firm i 's net demand for permits is $(r_{i,u} - r^T)q_{i,u}$, with negative values signifying a supply of permits. If the price of permits under an Intensity Target plan is $p_{IT,u}$, firm i 's profit maximization problem is

$$\max_{\{r_{i,u}, q_{i,u}\}} E_u[\pi_{i,u}^{IT}] = E_u \left[P_{i,u}(q_{i,u})q_{i,u} - c_i(r_{i,u})q_{i,u} - w_i q_{i,u} - p_{IT,u} q_{i,u} (r_{i,u} - r^T) \right] \quad (3.13)$$

The two first order conditions for an interior maximum are

$$-E_u[c'_i(r_{i,u}^{IT})] = E_u[p_{IT,u}] \quad (3.14)$$

if $q_{i,u}$ is greater than zero, and

$$E_u[P'_{i,u}(q_{i,u}^{IT})q_{i,u}^{IT} + P_{i,u}(q_{i,u}^{IT})] = E_u[c_i(r_{i,u}^{IT}) + w_i + r_{i,u}^{IT}p_{IT,u} - r^T p_{IT,u}] \quad (3.15)$$

Equation 3.14 is the usual efficient abatement condition that defines each $r_{i,u}^{IT}$. Equation 3.15 is the usual zero marginal profit condition which determines $q_{i,u}^{IT}$. If the output margin first order condition from 3.15 is compared with that under Cap and Trade in 3.12 one notices that the intensity Target equation contains an extra negative cost term associated with the Intensity Target itself. This term is due to the fact that the relative nature of the Intensity Target acts as an output and emission subsidy not found in a Cap and Trade system. Ultimately this is the source of inefficiency that authors in the literature described above attribute to Intensity Target emissions trading systems. Assume that the regulator sets the emission intensity target such that the expected aggregate emissions are the same as the optimum level found in the

SPM solution, so that

$$E_u \left[\sum_{i=1}^N r_{i,u}^* q_{i,u}^* \right] = E_u \left[\sum_{i=1}^N r_{i,u}^{CT} q_{i,u}^{CT} \right] = E \left[\sum_{i=1}^N r_{i,u}^{IT} q_{i,u}^{IT} \right] = \Omega^*, \quad (3.16)$$

where the emission intensity target is binding. If the net demand for permits in equilibrium is zero then Equation 3.3 holds, which implies that

$$\sum_{i=1}^N r_{i,u=0}^{IT} q_{i,u=0}^{IT} - \sum_{i=1}^N r^T q_{i,u=0}^{IT} = \sum_{i=1}^N r_{i,u=1}^{IT} q_{i,u=1}^{IT} - \sum_{i=1}^N r^T q_{i,u=1}^{IT} = 0 \quad (3.17)$$

Because firms make decisions about $r_{i,u}$ prior to knowing the state of u and this level is determined by Equation 3.14, it must be the case that $r_{i,u}^{IT} = r_i^{IT}$ in every output demand state, so

$$-c_i(r_{i,u}^{IT}) = -c_i(r_i^{IT}) = E_u[p_{IT,u}] = p_{IT}, \quad (3.18)$$

which implies that in equilibrium firms choose the same intensity level in every decision period. Equation 3.17 thus becomes

$$\begin{aligned} \sum_{i=1}^N r_i^{IT} q_{i,u=0}^{IT} - \sum_{i=1}^N r^T q_{i,u=0}^{IT} &= \sum_{i=1}^N r_i^{IT} q_{i,u=1}^{IT} - \sum_{i=1}^N r^T q_{i,u=1}^{IT} = 0 \\ \sum_{i=1}^N (r_i^{IT} - r^T) q_{i,u=0}^{IT} &= \sum_{i=1}^N (r_i^{IT} - r^T) q_{i,u=1}^{IT} = 0, \end{aligned} \quad (3.19)$$

implying that individual firms adjust only their output across states in order to clear the permit market. As well, Equation 3.15 becomes:

$$E_u \left[P_{i,u}(q_{i,u}^{IT}) \right] = c_i(r_i^{IT}) + w_i r_i^{IT} p_{IT} - r^T p_{IT}, \quad (3.20)$$

which means that

$$r_i^{IT} - r^T = \frac{E_u \left[P_{i,u}(q_{i,u}^{IT}) \right] - c_i(r_i^{IT}) - w_i}{p_{IT}} \quad (3.21)$$

This means that the difference between the intensity level selected by firm i and the target intensity level is equivalent to the marginal net benefit of output per dollar spent on emissions permits. For dirty firms the gains from production outweigh the costs, and so an intensity above the target level will be chosen. For clean firms the gains from production are less than the gains from selling permits and so intensities lower than the target will be chosen. Under Cap and Trade regulation without banking a similar equality holds:

$$r_i^{CT} = \frac{E_u \left[P_{i,u}(q_{i,u}^{CT}) \right] - c_i(r_i^{CT}) - w_i}{p_{CT}} \quad (3.22)$$

and implies that under Cap and Trade firms select an intensity level equal to the ratio of net benefits from output to the net costs of permits. From these two equations it is clear to see that if firms chose the same intensities and output quantities under both Cap and Trade and Intensity Targets, then the price of permits under Intensity Targets would be relatively higher than the price under Cap and Trade. According to the first order condition this difference in permit prices would cause firms under Intensity Targets to choose different, lower and cleaner, emissions intensity than they would under a Cap and Trade system.

3.5 Experimental parameterization and procedures

The experimental environment was based on a combination of the environments of Buckley et al. (2014) and Stranlund et al. (2014) with a few modifications. The environment is framed as a production decision in which firms require inputs to produce outputs, choosing the input ratio of their firm, the amount of output to produce and how many inputs to buy and sell in a double auction.

3.5.1 Parameterization

Each round, participants first make their emission intensity decisions followed by 180 seconds⁶ to produce output and buy or sell emission permits in a double auction. Each group of eight participants contains two types of equally represented 'clean' and 'dirty' firms. Firm types were randomly assigned and fixed for the duration of the experiment. All subjects began with an initial endowment of 250 lab dollars (L\$) in period 1. In the Cap and Trade case an endowment of permits was allocated to each firm at the start of each period. Uncertainty was defined by random shocks to the net marginal revenues (NMR) (where net marginal revenues are net marginal profits excluding permit costs) resulting in two NMR schedules, a low earnings schedule (L) and a high earnings schedule (H), for each intensity level, for each firm type. When firms cannot bank permits, these NMRs are the individual single-period permit demand schedules, which means they can be interpreted as marginal abatement costs (MAC) functions and equivalently, as permit demand schedules. These schedules are presented in Tables 3.2 and 3.3. The permit demand schedules for each firm type and intensity level are parallel, with the H values being L\$40 higher than the L values. Comparing the NMRs across firm types for a fixed level of intensity produces the output margin, which is modeled similar to Stranlund et al. (2014). Comparing the NMRs for a particular output level and firm type produces the intensity margin, which is modeled similar to Buckley et al. (2014). The work of both sets of authors is incorporated into this reduced form.

The general form for the average abatement cost function associated with

⁶Periods 1-6 lasted 180 seconds, the remaining periods lasted 120 seconds.

choosing different levels of emissions intensity is assumed to be

$$c_i(r_{i,u}) = u_{0,i} + (u_{1,i} - u_{0,i}) \left[\frac{r_{max} - r_{i,u}}{r_{max}} \right]^\alpha, \quad (3.23)$$

where $u_{0,i}$ and $u_{1,i}$ are firm-specific constants which define the bounds of the function and α is a convexity parameter which is the same for all firms. $r_{i,u}$ are integer intensity levels ranging from 0 to 5, with r_{max} being 5. This cost structure ensures that the marginal abatement cost curves for each firm are downward sloping and convex.

The experiment is implemented by presenting participants with NMR schedules. This formulation is consistent with an industry composed of firms with monopoly power in their respective output markets. Thus firms may face a different price for their output and different constant marginal costs of production, but the difference:

$$P_{i,u}(q_{i,u})q_{i,u} - c_i(r_{i,u})q_{i,u} - w_i q_{i,u} \quad (3.24)$$

is explicitly defined for each firm in each output demand state.

Apart from the demand state, firm costs differ in two dimensions: emission intensity and output. Firms choosing dirtier technologies embodied by higher emission intensities experienced lower cost (as given by Equation 3.23), and vice versa for choosing cleaner technologies. In addition, NMRs fell by L\$20 for each extra unit of output produced. This cost structure identified a unique profit maximizing emission intensity and output level for each subject each period. Participants knew which demand schedule they faced at the start of each trading period, following selection of an emission intensity. The complete permit demand schedules for each intensity level and output demand state were presented to subjects at the time of choosing their emission intensity level, along with a calculator with which they could test the impact

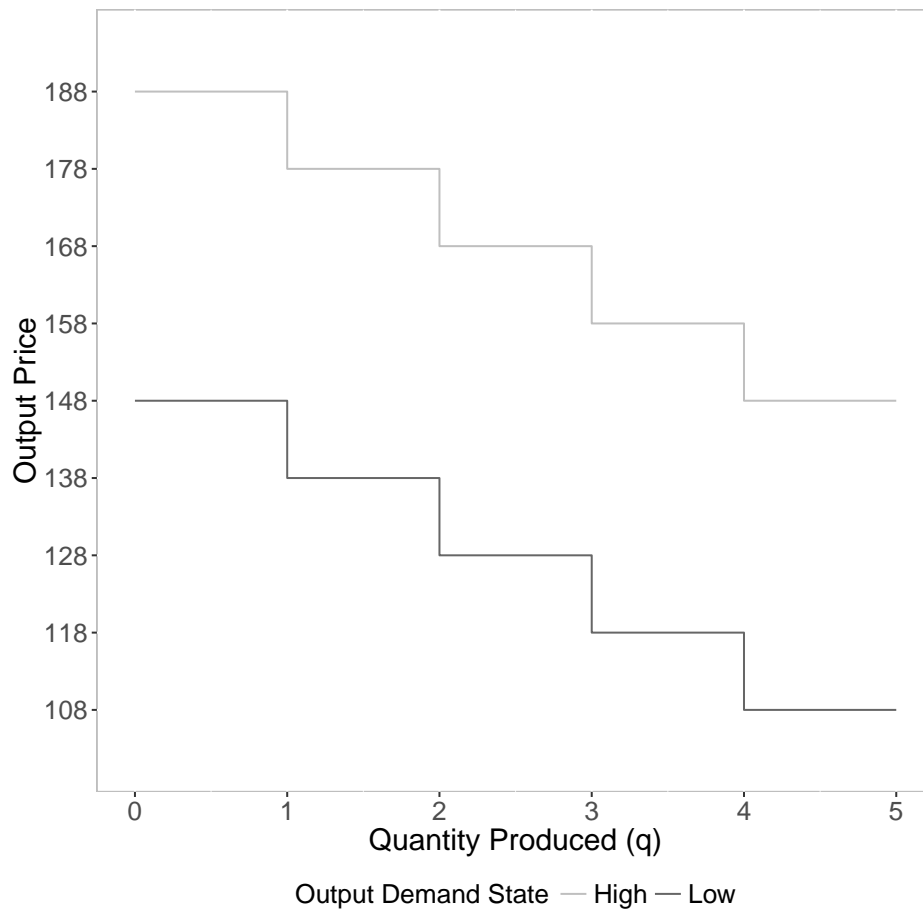


Figure 3.1: Implied output demand schedules.

of different permit prices on the H and L schedules for a particular intensity level.

The values contained in the NMR schedules in Tables 3.2 and 3.3 were generated using the equation $MR_u - c_i(r_{i,u})$. Equation 3.23 defines $c_i(r_{i,u})$. Table 3.1 identifies the parameters used to generate each value. In order to generate prices used in the calculation of consumer surplus the average marginal revenues (MR_u) for each unit were used. For example, if a firm produced 2 units of output and demand for output was high, then the average marginal revenue is $\frac{188+168}{2} = 178$, and this is the price per unit of output. The implied output demand schedules are presented in Figure 3.1.

The demand schedules implied by the NMR schedules that form the basis of

q	MR		u_0		u_1		α	w_i
	Low Demand	High Demand	Clean	Dirty	Clean	Dirty		
1	148	188	88	74	182	380	3	0
2	128	168	88	74	182	380	3	0
3	108	148	88	74	182	380	3	0
4	88	128	88	74	182	380	3	0
5	68	108	88	74	182	380	3	0

Table 3.1: Parameters used to generate net marginal revenue schedules.

the experiment are presented in Figure 3.1. The continuous form of Equation 3.24 was converted into a discrete representation for the purposes of the experiment and the integer values implemented. For each level of intensity, an additional unit of output decreases the NMR by L\$20.

Unit	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1	
	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH
1	-34	6	12	52	40	80	54	94	59	99	60	100
2	-54	-14	-8	32	20	60	34	74	39	79	40	80
3	-74	-34	-28	12	0	40	14	54	19	59	20	60
4	-94	-54	-48	-8	-20	20	-6	34	-1	39	0	40
5	-114	-74	-68	-28	-40	0	-26	14	-21	19	-20	20

Subjects 1, 3, 7, 8. All values rounded to the nearest integer. LOW and HIGH columns refer to the output demand state.

Table 3.2: Net marginal revenue of clean firms.

Unit	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1	
	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH
1	-232	-192	-83	-43	8	48	54	94	72	112	74	114
2	-252	-212	-103	-63	-12	28	34	74	52	92	54	94
3	-272	-232	-123	-83	-32	8	14	54	32	72	34	74
4	-292	-252	-143	-103	-52	-12	-6	34	12	52	14	54
5	-312	-272	-163	-123	-72	-32	-26	14	-8	32	-6	34

Subjects 2, 4, 5, 6. All values rounded to the nearest integer. LOW and HIGH columns refer to the output demand state

Table 3.3: Net marginal revenue of dirty firms.

In the Cap and Trade treatments an optimal amount of 48 permits were supplied to the market in each period. Firm types and permit endowments were fixed for the duration of the experiment. Following procedures similar to

those of Stranlund et al. (2014), permit endowment allocations were arbitrarily assigned within each firm type, as shown in Table 3.5, and chosen simply to provide trading opportunities for subjects in each period. With an equal assignment of permits few trades would take place.⁷ In each Intensity Target treatment there was a global emission intensity target r^T of 2 units of emissions per unit of output, the target predicted to result in an aggregate emission level identical to that under the Cap and Trade treatment.

In total 6 sessions involving Cap and Trade and 12 sessions involving Intensity Targets were run. More Intensity Target sessions were run because this system is least understood, with most of the current literature focused solely on Cap and Trade emission trading. In every period states H and L were equally likely and determined by sorting a random draw from a uniform distribution on $[0,1]$ as L if the random number was less than 0.5 and H otherwise. Every sequence lasted at least 10 periods with certainty and this was common knowledge in the experiment. In order to discourage end game effects, after round 10 and every subsequent period there was a $\frac{1}{6}$ chance that the session would end. This was determined by a random draw from the sequence (1,2,3,4,5,6), where 6 indicated that the session ended. Because only a relatively small number of sessions of each treatment were being run and the total number of permutations for different play lengths would be very large should independent and identically distributed draws be run for every session, prior to the start of the experiment 2 sequences of random demand shocks were generated. The result was Sequence 1 extending for 12 periods and Sequence 2 for 14 periods. Participants were told that the random numbers used to select demand states in their session were pre-determined and programmed into the experimental software. In pilot studies subjects

⁷This type of allocation of permits would not be necessary to generate trading in the field because firms have far more heterogeneous cost structures than those in our experiment.

expressed no apprehension involving the predetermined random values we used to ensure replication. In each treatment Sequence 2 was used at least once and Sequence 1 at least twice, Table 3.4 presents each sequence of output demand states used in the experiment.

Period	Sequence 1	Sequence 2
Practice 4	LOW	LOW
Practice 5	HIGH	HIGH
Practice 6	LOW	LOW
1	HIGH	LOW
2	LOW	HIGH
3	HIGH	HIGH
4	HIGH	LOW
5	LOW	LOW
6	LOW	HIGH
7	LOW	HIGH
8	LOW	HIGH
9	HIGH	LOW
10	HIGH	LOW
11	LOW	HIGH
12	HIGH	LOW
13	-	HIGH
14	-	LOW

Table 3.4: Demand sequences.

The equilibrium permit demand schedules presented in Figures 3.2 and 3.3 trace the the aggregate demand for permits in the Cap and Trade environment on the intensity margin, assuming output is held fixed at the expected level of output, and on the output margin assuming firms select their optimal equilibrium emission intensity levels.

The equilibrium permit price under Cap and Trade for the model and parameter choices stated above is L\$16 , and is L\$28 under the Intensity Target treatment. At theses predicted prices firms are maximizing profits on both the output and emissions intensity margins and the permit market will clear. Turning to Figure 3.2, setting the aggregate cap equal to 48 permits in each period leads to three expected price tunnels, depending on output de-

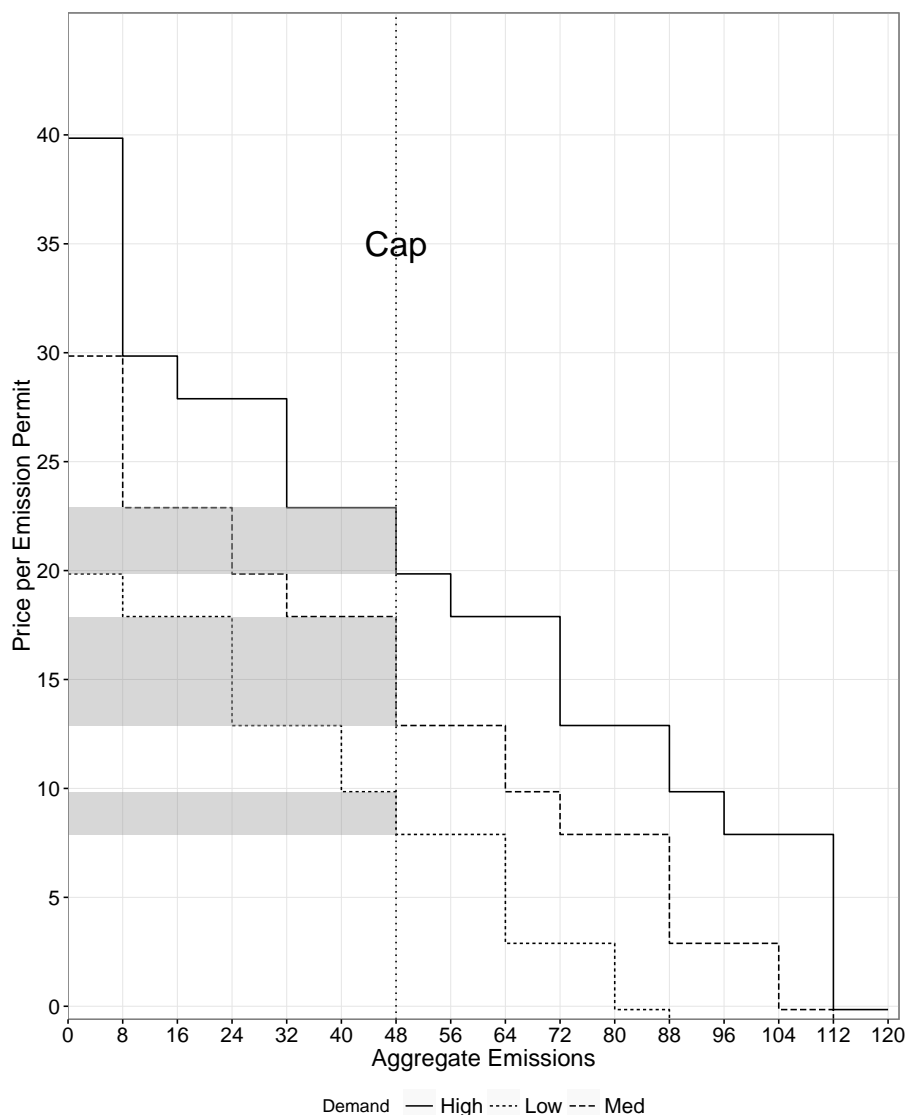


Figure 3.2: Cap and Trade emission permit demand schedules: Output margin.

demand and whether banking of permits is allowed. With high demand for output and a cap of 48 permit prices are expected to be within L\$20 and L\$23 if banking is not allowed. With low demand for output and a cap of 48 permit prices are expected to be between L\$8 and L\$10 when banking is not allowed. The expected permit price when high and low states are equally represented and banking is allowed is thus between L\$13 and L\$18, a range that contains our L\$16 equilibrium permit price prediction.

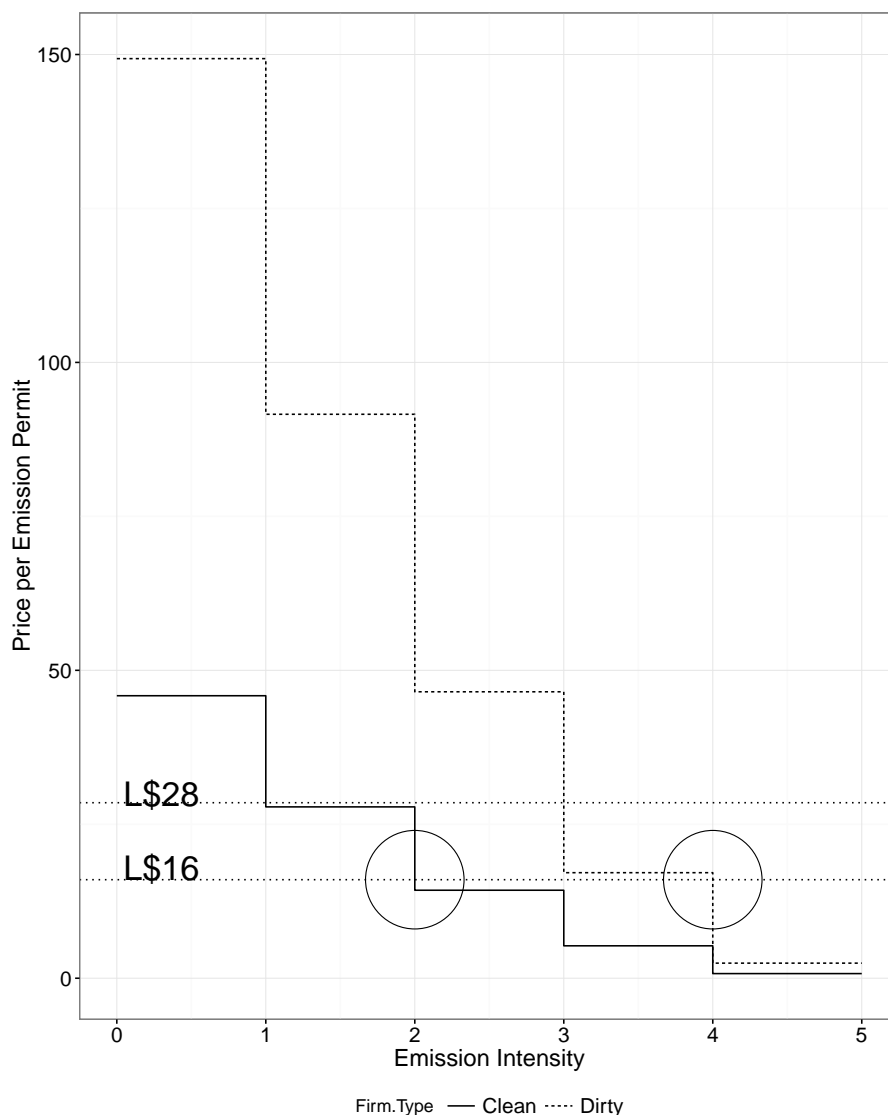


Figure 3.3: Cap and Trade marginal abatement cost curves: Intensity margin.

Under Cap and Trade without Banking the total supply of permits in the market in any period is equal to the cap. Since intensity decisions are made prior to having information about demand states we take the expected price of L\$16 and find the optimal emission intensities for each firm type in Figure 3.3. At a price of L\$16 it is expected that clean firms will select an intensity of 2 units of emissions per unit of output and dirty firms an intensity of 4. This intensity selection is independent of the demand state. If, however, firms can bank permits a slightly different process takes place. Turning again to Fig-

ure 3.2, when output demand is low permit prices are expected to be between L\$8 and L\$10. At this price the expected demand is 72 permits. When output demand is high, permit prices are expected to be between L\$20 and L\$23, and the expected demand for permits is 24. Since firms can bank permits, they can buy permits when prices are low reserving these permits for use when permit prices are otherwise high, adapting to demand conditions. The expected total emissions when demand states are equally balanced is 48 and output should vary between high and low output. Again, the expected price of permits is L\$16 and so the optimal firm intensity choices are the same as in the case without banking and follow the same process of selection.

Under Intensity Targets the price of permits is expected to be higher, at L\$28 rather than the L\$16 predicted under Cap and Trade, because the intensity target acts as a subsidy on output, increasing emissions and therefore increasing permit prices. At a price of L\$28 clean firms under Intensity Targets are expected to select an intensity of 1 and dirty firms an intensity of 3, also independent of demand for output. Table 3.5 presents the firm specific parameters which characterize the equilibrium in which emissions are predicted to be identical under both Cap and Trade and Intensity Target treatments. Notice that, due to the inherent subsidy on output, firms under Intensity Target regulation must set lower and more costly average emission intensity levels to achieve the same predicted aggregate emissions as under a Cap and Trade plan. While Intensity Targets do have this drawback, strategic banking is not required by profit maximizing firms facing Intensity Targets, while those facing Cap and Trade must do so.

Firm Type	Optimal Emission Intensity (CT)	Optimal Emission Intensity (IT)	Optimal Output Low Demand (CT)	Optimal Output Low Demand (IT)	Optimal Output High Demand (CT)	Optimal Output High Demand (IT)	Permit Endowment Each Period (CT)	Intensity Target (IT)
Clean	2	1	1	2	3	4	1 or 6	2
Dirty	4	3	1	2	3	4	11 or 6	2

Table 3.5: Experiment parameters and predictions.

3.5.2 Procedures

We ran 18 laboratory sessions at the McMaster Experimental Economics Laboratory. Three sessions were run with Cap and Trade with Banking allowed (CTB), three sessions were run with Cap and Trade with No Banking allowed (CTNB), six sessions were run with Intensity Targets with Banking allowed (ITB) and six sessions were run with Intensity Targets with No Banking allowed (ITNB). Each session involved eight subjects for a total of 108 participants. Subjects were recruited from the general population of undergraduates at McMaster University. Each session of the experiment lasted approximately 2 hours and consisted of two sets of unpaid practice periods and either 12 or 14 paid periods. During the first set of three practice periods students received instructions and practiced producing output and making trades in a double auction environment under a set of parameters different from those applied in the paid portion of the experiment. In the second set of three practice periods students faced an additional stage in which they chose an 'input intensity ratio' and faced uncertainty about the demand for output. After each member of the group made an intensity decision, the group proceeded simultaneously to the same production and trading screen as in the first set of practice periods where the demand for output was realized. The parameters of this second practice set were different from the first practice set and also from the parameters used in the paid portion of the experiment. Prior to beginning the practice periods, participants completed the risk assessment device described in Holt and Laury (2002). At the end of the experiment participants were informed of their results and paid privately in cash. Subjects earned between C\$ 15.78 and C\$ 44.42, not including the C\$5 show up fee or earnings from the risk assessment task. The software implementation of the environment was programmed using the z-tree software package (Fis-

chbacher, 2007).

3.5.2.1 Efficiency

We define Efficiency as in Buckley et al. (2014), where

$$\text{Efficiency} = \frac{\text{Observed Total Surplus}}{\text{Maximum Surplus Available}} \quad (3.25)$$

The total surplus is composed of the consumer's surplus and producer's surplus less environmental damage. Our specification of the output market enables the calculation of consumer surplus for firms operating as monopolists in distinct output markets.⁸

$$\text{Total Surplus} = \text{Producer Surplus} + \text{Consumer Surplus} - \text{Damages} \quad (3.26)$$

Two ways of defining the denominator of the efficiency equation 3.25, maximum total surplus available, are considered. The first way is defined as 'within' efficiency because the maximum surplus available within a particular treatment is used. The second way, termed the 'overall' efficiency, sets the maximum total surplus as the maximum surplus available across any of the treatments. Because the ultimate aim of a permit market is to maximize total surplus, this provides a comparison across treatments relative to the treatment with the greatest total surplus.

⁸For example, if a firm produced 2 units of output and demand for output was high, then the average marginal revenue, from Table 3.1 is $\frac{188+168}{2} = 178$, and this is the price per unit of output used to calculate consumer surplus.

3.5.3 Summary of parameterization and procedures

Relative to Cap and Trade, Intensity Targets should be a less efficient means of controlling emissions. This is because intensity targets act as a subsidy on output, increasing total emissions, and imposing higher costs on firms to meet an equivalent expected level of emissions.

Our experimental design allows us to test hypotheses concerning the relative performance of Cap and Trade and Intensity target regimes and banking provisions. For each of the Cap and Trade and Intensity Target treatments we run a banking and no-banking variant. In order to maximize total surplus under Cap and Trade, banking is necessary across periods, while for Intensity Targets this is not the case because in equilibrium permit demand equals permit supply in each period. This is an attractive simplification to the administration of an Intensity Target plan. However, there may be other behavioural forces at work in a setting of uncertain demand for output. Firms may choose to bank permits in response to uncertainty even when it is not optimal to do so, as was found by Buckley et al. (2014). We explicitly provide an environment in which permit banking is not necessary in equilibrium but where subjects can bank permits should they choose to, and compare the outcome to an environment in which permit banking is not possible.

3.6 Predictions and results

We provide results on per-period aggregate emissions and output, and average firm intensity choices, by treatment and firm type. As well, we examine efficiency and prices of emission permits, examining prices in each output demand state. Analysis was undertaken in R and data was imported using the zTree to R package offered by Kirchkamp (2013).

The experiment was parameterized such that aggregate expected emissions

were equivalent in all treatments, given a balanced number of high and low demand states. We encountered evidence of learning effects in the first 3 periods which dissipated by period 4. Periods 1-3 were thus dropped from the analysis. Demand sequences 1 and 2 were of length 12 and 14 periods respectively. We chose to use periods 4-12 for analysis, as this equated the total number of observations in each sequence. In addition, the number of High and Low demand states is the same in each sequence over periods 4-12, and so the expected results are the same regardless of the sequence. Box-plots of the mean results for periods in blocks (periods 1-12, 4-12, 4-9, 1-3,10, and 10-12) also indicated a potential for end-game effects for some outcomes in some treatments but the overall effect of excluding periods 10-12 (leaving only periods 4-9 in the analysis) was not significantly different from the results of periods 4-12 in the majority of cases, and using periods 4-12 resulted in the same number of high and low demand states in each sequence. In periods 4-12 there were 5 occurrences of the low demand state and 4 occurrences of the high demand state. The predicted per-period aggregate emission levels are calculated using the predicted optimal intensity and output levels in each state. Because in Cap and Trade with Banking, as well as in both Intensity Target treatments the optimal predicted output values differ depending on the particular demand state, the mean of the predicted aggregate values differs slightly from the parameterized value due to the fact that each session contained more low demand states than high demand.

The step functions which characterize the profit functions make the equilibria difficult to solve for directly, therefore a numerical search approach was adopted. Under Intensity Targets if all firms select the target level of emissions then no emission permits are generated or demanded. In this situation the marginal net revenues of clean firms are positive up to the third unit of output in the low demand state and five units of output in the high demand

state. Total earnings in each state are L\$60 and L\$200. For dirty firms the low and high state output quantities are 1 and 3 and total earnings L\$8 and L\$84. However, if there is to be trade in permits then it must be the case that one firm type supplies permits while the other type requires permits. Permits are supplied when firms select intensities below the target level of 2 units of emissions per unit of output. Tables 3.2 and 3.3 demonstrate that for intensity levels below the target clean firms have lower costs of supplying permits than dirty firms, strictly dominating dirty firms in terms of supplying permits. The size of the permit market is thus determined by the supply decisions of clean firms. This is because dirty firms are required to hold sufficient permits at the time of production, and therefore must buy permits from clean firms prior to production. If clean firms choose an intensity of 0 then 2 emission permits are generated for each unit of output produced by a clean firm. Under low output demand conditions the first unit of output produced at this intensity level generates a loss of L\$34. In order to produce this output the clean firm must therefore expect to receive at least L\$17 for each permit generated.

Demand for emission permits is driven by dirty firms choosing intensity levels above the target. At every intensity level dirty firms are willing to buy permits at prices higher than the price required by clean firms. However, in equilibrium the intensity level chosen by dirty firms will result in net demand for permits which is equal to the net supply of permits. Choosing intensity levels which result in lower or higher net demand for permits than supplied by the clean firm will result in permit prices which are lower or higher than market clearing prices, providing incentive for dirty firms to choose an intensity level which leads to clearing of the permit market.

In order to satisfy permit market clearing, if clean firms choose an intensity of 0, then dirty firms should choose an intensity level of 4. Thus clean firms

supply 2 permits for each unit of output produced and dirty firms demand 2 emission permits for each unit of output produced. Dirty firms are willing to pay up to L\$36 for each permit if one unit of output is produced. Clean firms are willing to accept any amount greater than L\$17 for each permit. One unit of output will therefore be produced. For the second unit of output Clean firms will accept any amount greater or equal to L\$27 for each permit and dirty firms are willing to pay any amount less than or equal to L\$26. Since the dirty firms are unwilling to pay the amount required by clean firms for the second unit of output, in equilibrium this second unit of output will not be produced at this intensity level.

Clean firms choosing an emission intensity of 0 is not an equilibrium. This is because clean firms can produce one unit of output at a lower cost if they choose a slightly dirtier emission technology of 1 unit of emissions per unit of output. At this intensity level under low demand clean firms earn a positive net revenue regardless of the price of permits. If dirty firms continue to choose an intensity level of 4 while clean firms choose the intensity of 1 then an excess demand for permits will result, driving up permit prices and providing an incentive to dirty firms to choose the slightly cleaner emission intensity level of 3 units of emissions per unit of output. Under low output demand and at the intensity level of 3, dirty firms are willing to pay any amount less or equal to L\$54 for each permit for the first unit of output. Since clean firms are willing to supply that permit at any price, it will be available on the permit market.

For the second unit of output, clean firms are willing to accept any amount greater or equal to L\$8 for each permit. Dirty firms are willing to pay any amount less or equal to L\$34 and so a second unit will be produced. Under low demand, a third unit will not be produced in equilibrium. This is because clean firms are willing to produce if the price of permits will be greater or

equal to L\$28, but dirty firms are willing to pay up to only L\$14. The equilibrium price of permits with these particular intensity choices and this level of output occurs at permit prices of L\$28.⁹

Clean firms are therefore expected to choose an intensity of 1 unit of emission per unit of output, and dirty firms an intensity of 3 units per unit of output.

This choice is independent of the output demand state. Once output demand is known, under low demand clean firms will produce 2 units of output for a total return L\$4 and generate 2 emission permits earning a price of L\$28 for each permit, for a total profit of L\$60 in the low output demand state. Likewise, dirty firms will produce 2 units of output earning L\$88 and demand 2 emission permits, paying L\$28 per unit, earning total profits of L\$32. Under high output demand clean firms produce 4 units of output, generate 4 permits and earn total profits of L\$200. Dirty firms will produce 4 units of output and demand 4 emission permits, earning total profits of L\$144.

In equilibrium under low output demand, each clean firm generates 2 units of emissions in total (intensity of 1 for 2 units of output), and each dirty firm generates 6 units of emissions (intensity of 3 for 2 units of output). Under high output demand clean firms each generate 4 units of emissions and dirty firms 12 units of emissions. At the industry level total emissions are thus 32 in the low state and 64 in the high state. The expected level of emissions is thus 48, which is the level of the aggregate cap under Cap and Trade.

In order to generate predictions specific to the experiment, the expected values of emissions, output and intensities are generated using the exact demand sequence implemented in the experiment. In each sequence for periods 4-12 there are 5 periods of low demand and 4 periods of high demand.

This results in a predicted aggregate emissions level of approximately 46

⁹Taking the continuous functions into the discrete setting, values which were presented to subjects were rounded. The exact price in equilibrium is within L\$ [27.824, 28.128]. Subjects could enter only integer valued prices in the experiment.

units in each period under Intensity Targets. Under Cap and Trade without Banking subjects are limited by the available cap in each period, and are expected to produce the amount of the cap in each period. With banking however, subjects are expected to restrict output in low demand states resulting in aggregate emissions of 24, and expand output in high demand states resulting in aggregate emissions of 72 in high demand states. The expected value is thus 48 emissions, but extended over a demand sequence with 5 low demand states and 4 high demand states leads to an expected value of approximately 45 total units of emissions per period for the experiment.

Table 3.6 summarizes the equilibrium predictions and per-period, per-session average results under each trading mechanism, pooled over firm types.

	Emissions		Output		Intensity		Permit Prices	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs
^a CT B	45	50.67 ^{*,b,c,d}	15	14.37 ^{c,d}	3	3.4 ^{*,c,d}	16	16.59 ^{c,d}
^b CT NB	48	44.3 ^{*,a,d}	16	13.74 ^{*,c,d}	3	3.11 ^{c,d}	15	17.19 ^{c,d}
^c IT B	46	39.3 ^{*,a,d}	23	19.24 ^{*,a,b}	2	2.31 ^{*,a,b}	28	38.33 ^{*,a,b}
^d IT NB	46	33.54 ^{*,a,b,c}	23	17.13 ^{*,a,b}	2	2.24 ^{*,a,b}	28	40.02 ^{*,a,b}

Differences from predictions are estimated using box-plots comparing session averages for periods 4-12 to the predicted values because with only three per-session observations in each Cap and Trade treatment no statistical test can detect a significant difference. * Indicates a clear difference between the predicted value and the observed results using a box-plot. See Appendix Section 3.A, Figure 3.A.1 for the plots.

^{a,b,c,d} refer to a significant difference at the 10% level of the treatment with the super-scripted treatment (a: Cap and Trade with Banking, b: Cap and Trade without Banking and c: Intensity Target with Banking, d: Intensity Target without Banking) using a two-sided two-sample (un-paired) Wilcoxon rank sum test with continuity correction.

Table 3.6: Summary of predictions and observed results by treatment pooled across firm types.

3.6.1 Results by treatment

Table 3.6 summarizes the equilibrium predictions and per-period, per-session average results under each trading mechanism, pooled over firm types.¹⁰ Un-

¹⁰Because emission intensities were integer choices we considered the use of the mode but found that in this case average intensity values capture the non-uniform nature of intensity

der Intensity Targets aggregate emissions are significantly lower than the prediction in both banking and no banking treatments. Firms tend to choose dirtier production technology than expected, and produce less output than expected. Permit prices in the Intensity Targets treatments are significantly higher than predicted, which may be a key influence in the production decisions of dirty firms, who must buy permits from clean firms in order to produce output. Section 3.6.3 will investigate decisions made by each firm type. Across treatments, aggregate emissions under Cap and Trade with Banking are significantly higher than emissions in all other treatments, and Intensity Targets without Banking significantly lower. Output is significantly higher in Cap and Trade treatments and lower in Intensity Target treatments, compared to their respective predictions, but not different across banking and no-banking treatments for Cap and Trade. Average intensity levels differ across permit trading schemes, but not across banking treatments. Average permit prices differ across Cap and Trade and Intensity Target trading schemes, but not across banking treatments within each trading scheme.

Overall, the results suggest that although Intensity Targets are expected to result in higher costs (due to lower emission intensities) for the same level of emissions as the Cap and Trade regulation, this may not be borne out in practice to the extent predicted. For a system calibrated to produce the same expected level of aggregate emissions as Cap and Trade regulation, Intensity Target regulation consistently produces lower levels of emissions, while at the same time providing consumers with more output than under Cap and Trade. In addition, when Intensity Targets are implemented, banking does not significantly affect permit prices, average intensity levels or output. However, total emissions are higher when banking is permitted in both permit trading schemes.

choices better than modal values.

3.6.2 Efficiency

Table 3.7 presents two measures of market efficiency. Efficiency is defined in Section 3.5 as the observed total surplus divided by the maximum total surplus. ‘Within’ efficiency compares observed total surplus to the maximum total surplus attainable within a treatment. ‘Overall’ efficiency compares observed total surplus to the maximum total surplus of all treatments.

	Producer Surplus		Consumer Surplus		Damage		Total Surplus		Efficiency (%)	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Within	Overall
^a CT B	1060	979*	107	110	725	811*	441	279*	63	63 ^{b,c,d}
^b CT NB	1020	839*	80	92	768	709*	332	223*	67	50 ^a
^c IT B	816	626*	258	226	740	629*	334	223	67	51 ^a
^d IT NB	816	559*	258	187*	740	537*	334	209*	63	47 ^a

Results rounded to the nearest integer.

Differences from predictions are estimated using box-plots comparing session averages for periods 4-12 to the predicted values because with only three per-session observations in each Cap and Trade treatment no statistical test can detect a significant difference. * Indicates a clear difference between the predicted value and the observed results using a box-plot. See Appendix Section 3.A, Figure 3.A.2 for the plots.

^{a,b,c,d} refer to a significant difference at the 10% level of the treatment with the super-scripted treatment (a: Cap and Trade with Banking, b: Cap and Trade without Banking and c: Intensity Target with Banking, d: Intensity Target without Banking) using a two-sided two-sample (un-paired) Wilcoxon rank sum test with continuity correction.

Table 3.7: Summary of efficiency measure predictions and results by treatment pooled across firm types.

The ‘within’ efficiency is highest in the case of Cap and Trade without Banking and lowest under Intensity Targets with Banking. Relative to the treatment which produces the highest total surplus, the ‘overall’ efficiency is highest under Cap and Trade with Banking and significantly higher than the efficiencies observed under Intensity Targets. This result is consistent with the findings of Fischer (2001), who suggested that Intensity Targets represent an efficiency loss over optimal emissions pricing, and consistent with the findings of Buckley et al. (2014), although both of these papers assume that output demand is certain, unlike the experiment presented here. What is found in the current experiment is that the within efficiencies are not significantly

different from each other, suggesting that all emission permit regulations tended to produce similar results in terms of achieving the total surplus expected.

Consumer and producer surplus components were calculated using the MR values from Table 3.1 and the cost function in Equation 3.23, according to the output and intensity choices made by participants. Damage was calculated using the assumption of a flat damage function with a constant marginal damage of L\$16 (the optimal permit price under Cap and Trade). While the assumption of constant marginal damage costs is consistent for a stock pollutant like greenhouse gases, assuming increasing marginal damages would result in increased efficiency for the Intensity Targets treatment due to their low emission levels. In this way the overall efficiencies in Table 3.7 should be viewed as a conservative estimate for the Intensity Target treatment.

3.6.3 Results by firm type

Table 3.8 provides the results of per-period aggregate emissions, aggregate output and intensity choices by firm type. Under Cap and Trade with Banking emissions are consistent with predictions for clean firms, but not for dirty firms, who produce slightly more than predicted. Without banking the relationship flips, with clean firms producing a substantial amount less than predicted and dirty firms producing as predicted. In all Cap and Trade cases output differs from the predicted levels, with clean firms producing less and dirty firms more than predicted. The intensity choices for all Cap and Trade firms align with the predictions except for clean firms without banking who selected, on average, higher intensities than predicted. No detectable difference between firm types is identified using the usual test procedure, but the box-plots in Figure 3.A.3 suggest a substantial difference in emissions, output and intensities across firm types for the Cap and Trade treatments both with

and without banking.

	Emissions		Output		Intensity	
	Pred	Obs	Pred	Obs	Pred	Obs
CT B Clean	15	17.22 [†]	8	6.19 ^{*,†}	2	2.84 ^{*,†}
CT B Dirty	30	33.44 ^{*,†}	8	8.19 ^{*,†}	4	3.96 [†]
CT NB Clean	16	11.89 ^{*,†}	8	4.96 ^{*,†}	2	2.49 [†]
CT NB Dirty	32	32.41 [†]	8	8.78 ^{*,†}	4	3.72 [†]
IT B Clean	12	20.94 [*]	12	12.83 [†]	1	1.82 ^{*,†}
IT B Dirty	35	18.35 [*]	12	6.41 ^{*,†}	3	2.81 [†]
IT NB Clean	12	16.83 [*]	12	11.06 [†]	1	1.76 ^{*,†}
IT NB Dirty	35	16.7 [*]	12	6.07 ^{*,†}	3	2.72 [†]

* Indicates a clear difference between the predicted value and the observed results using a box-plot. See Appendix Section 3.A, Figure 3.A.3 for the plots.

† Indicates a clear difference between the observed results for the clean and dirty firms given a treatment using a box-plot. See Appendix Section 3.A, Figure 3.A.3 for the plots.

Table 3.8: Summary of predictions and results by treatment and firm type.

Under Intensity Targets with Banking the aggregate emissions of both clean and dirty firms are not significantly different from each other. This means that clean firms emit significantly greater emissions than predicted, while dirty firms emit significantly less than predicted. Output and emission intensity suggest that the equalized emission levels are driven by clean firms who produce significantly higher than predicted output and adopt significantly dirtier emission intensities than predicted. Similar results are found under Intensity Targets when there is no provision for emission permit banking. What these results indicate is that the incentives of the market structure under Intensity Targets are sufficiently strong that dirty firms do not produce output, and therefore emissions, at the levels predicted.

Overall it would appear that the incentives are such that clean firms are in a more desirable position under Intensity Targets. Clean firms have two areas in which they face potential gains - one by producing output, and therefore collecting emission permits, and again by selling emission permits to dirty firms. Under Intensity targets, the results of clean firms suggest that

they chose emission intensities which were dirtier than predicted, and therefore collected fewer emission permits than are predicted in equilibrium. With fewer permits available, permit prices were higher than expected and emissions of dirty firms lower than predicted. In response to the higher permit prices dirty firms choose cleaner emission intensities than predicted, which also contributed to the lower emissions results. The no-borrowing constraint in the permit market meant that the decreased supply of permits directly limited the ability of dirty firms to produce output because permits on hand were required prior to production. Clean firms thus determine the number of permits available and the prices of those permits.

3.6.4 Emission permit prices

Per-period average permit prices are unaffected by output demand and banking under Cap and Trade as shown in column CTB and CTNB of Table 3.9. However, as shown in the last two columns of Table 3.9, under Intensity Targets output demand significantly influences the average prices of emission permits, prices are higher when demand for output is higher, regardless of banking provisions. This result is consistent with the theoretical finding that in equilibrium banking is not required for optimization under Intensity Targets and that Intensity Target programs are more responsive to demand conditions.¹¹

¹¹There was insufficient information to assess the time paths of permit prices, aggregate emissions and the permit bank under Cap and Trade with Banking detailed in Stranlund et al. (2014).

Demand State	CTB		CTNB		ITB		ITNB	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs
Low Demand	9	15.73*	15.5	13.64	28	37.89*	28	36.86*
High Demand	21.5	17.66*	15.5	21.63	28	42.28*	28	43.97*

* Indicates a clear difference between the predicted value and the observed results using a box-plot. See Appendix Section 3.A, Figure 3.A.4 for the plot.

^D Indicates a clear difference between the observed results for each demand state given Cap and Trade or Intensity Targets using a box-plot. (None are detected). See Appendix Section 3.A, Figure 3.A.4 for the plot.

^B indicates a significant difference in results between banking treatment status at the 10% level using a paired Wilcoxon signed rank test with continuity correction (No significant differences are detected).

Table 3.9: Summary of mean permit prices by output demand state and treatment banking status, pooled across firm types.

3.6.4.1 Emissions and permit price volatility

A primary finding of Stranlund et al. (2014) was that banking and price controls resulted in less volatile permit prices than their baseline case of no policy, but that a decrease in price volatility came at the cost of more volatile emissions. Emissions volatility could be highly relevant to the specific type of pollutant for which the emissions market attempts to control. For stock pollutants volatility is less of a concern than reducing the total emissions produced, but for flow pollutants, higher levels imply greater damages and so decreasing both absolute amounts and controlling volatility is important. The theoretical model proposed in Section 3.4 reduced to one of static optimization in all cases except Cap and Trade with Banking, so expected emissions and permit prices should be constant over time in all programs except, potentially, under Cap and Trade when banking is permitted. Table 3.10 presents the mean variance of the per-session per-period emissions and mean permit prices. Mean volatility is also provided, and defined as in Stranlund et al. (2014):

$$\text{volatility} = |x_{g,t} - x_{g,t-1}|, \quad (3.27)$$

where $x_{g,t}$ is aggregate emissions or average per-period price of emission permits for a given session g in period t .

Table 3.10 highlights that Intensity Targets with Banking result in lower levels of emission variance than Cap and Trade with Banking. Only Intensity Targets without Banking result in significantly higher variance in permit prices. It is important to note that the variance and volatility values are dependent on the particular parameterization of the experiment and therefore not comparable across Cap and Trade and Intensity Target treatments, however the ranking of volatilities found by Stranlund et al. (2014) across banking treatments is replicated here. Banking results in significantly greater variance and volatility in emissions under both Cap and Trade and Intensity Target treatments. At the same time, this greater variance and volatility in emissions is associated with lower variance and volatility in permit prices, though the result is significant only for Intensity Targets.

	Emissions				Prices			
	Variance		Volatility		Variance		Volatility	
	Pred	Obs	Pred	Obs	Pred	Obs	Pred	Obs
^a CT B	576	695.46 ^b	24	24.04 ^b	16	26.05 ^d	0	3.55*
^b CT NB	0	9.56 ^a	0	2.46 ^{*,a}	43	34.75	6.5	5.51
^c IT B	256	253.02 ^a	16	17.54 ^a	0	37.19 ^d	0	5.65 ^{*,d}
^d IT NB	256	161.12 ^{*,a}	16	12.69 ^a	0	164.87 ^{*,c}	0	9.67 ^{*,c}

* Indicates a clear difference between the predicted value and the observed results using a box-plot. See Appendix Section 3.A, Figure 3.A.5 for the plots.

^{a,b,c,d} refer to a significant difference at the 10% level of the treatment with the super-scripted treatment (a: Cap and Trade with Banking, b: Cap and Trade without Banking and c: Intensity Target with Banking, d: Intensity Target without Banking) using a two-sided two-sample (un-paired) Wilcoxon rank sum test with continuity correction.

Table 3.10: Summary of mean variance and volatility by treatment, pooled across firm types.

3.6.5 Intensity choices and risk attitudes

Ben-David et al. (2000) suggest that risk attitude may play a role in intensity choices and compliance decisions in an environment in which these choices are irreversible. While we allow for reversible technology decisions by permitting the unrestricted choice of an emissions intensity, in our Intensity Target environment it is possible for participants to avoid exposure to emission permit market risk by choosing an intensity at the target level. In fact, we observed that clean firms choose on average higher intensity levels than predicted (i.e. choosing intensities closer to the target) even though clean firms stood to gain from clean technology by selling permits. We also observed that dirty firms tended towards lower intensity levels than predicted. If risk attitudes influence these decisions we would expect therefore that smaller absolute differences from the target to be correlated with greater levels of risk aversion.

We ran a well known risk assessment protocol designed by Holt and Laury (2002) at the start of each session to collect risk attitudes from participants. Under this protocol a series of 10 decisions between two lottery choices is required, with one lottery choice representing a less risky option. The number of less risky choices is used here as a measure of the level 'risk aversion' exhibited by each participant.

Throughout the experiment each participant played the role of a single firm type, and all periods had the same intensity target. The absolute value of the distance between the mean of all emission intensity decisions made by each subject and the intensity target represents our measure of 'closeness' to the intensity target. For each session, the correlation of this measure of closeness and the level of risk aversion exhibited by each participant is calculated using Kendall's τ statistic for correlation within data which are not necessarily

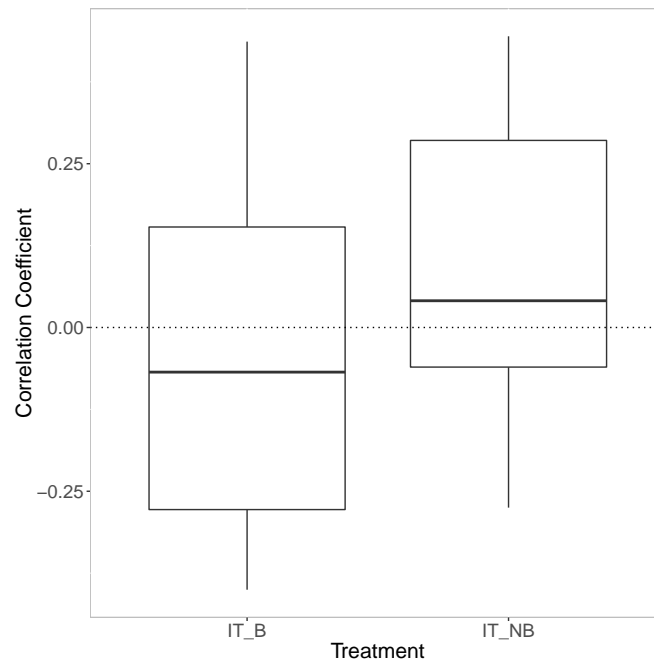


Figure 3.4: Correlation of ‘closeness’ to target and level of risk aversion.

normally distributed. The mean of these correlation values and the associated Wilcoxon test of the difference in these values from 0 are reported in Table 3.11. This table suggests that at the treatment level there is no evidence for a systematic relationship between risk attitude and choice of emission intensity. Figure 3.4 illustrates the distribution of correlation coefficients.¹² Testing by firm types does not reveal any additional insight.

Treatment	Mean Correlation	P-value
ITB	-0.04	0.8339
ITNB	0.09	0.4185

Average per-session Kendall’s τ correlation coefficients p-values.

Table 3.11: Correlation of ‘closeness’ to target and level of risk aversion.

Another means to assess the role of risk attitude in influencing participant behaviour is to investigate whether more risk averse subjects held larger banks of permits in an attempt to protect themselves from uncertainty. Figure 3.5

¹²A plot of the per session average difference from the intensity target and risk attitude level can be found in Appendix Section 3.A, Figure 3.A.6

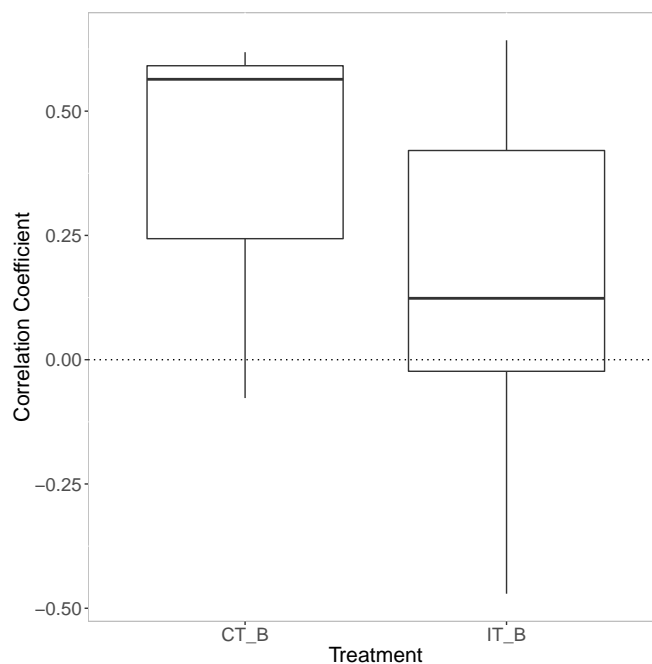
Kendall's τ correlation coefficient

Figure 3.5: Correlation of average permits banked and level of risk aversion.

shows the average per-session Kendall's τ correlation coefficients for treatments in which banking was allowed. Under Cap and Trade, banking was required in order to optimize and in these sessions there is a positive correlation between risk aversion and the average number of permits banked. The box-plot demonstrates that the interquartile range (covered by the extended bars) crosses zero and therefore this result is not significant.¹³ Under Intensity Targets, where banking was not required, a positive association between greater levels of risk aversion and the average permit bank is also observed, though this association is also not significant. Across treatments the results are not significantly different.

¹³Only the box-plots are shown here because the small number of per-session observations in the Cap and Trade treatment prohibits statistical testing.

3.7 Discussion and conclusions

This study compared Cap and Trade to Intensity Target emission permit trading programs in a setting in which firms choose both emission intensity and output levels and face stochastic shocks to output demand. In a controlled laboratory environment a market parameterized to produce the same level of expected emissions across trading programs resulted in significantly lower emissions and output than expected under Intensity Targets as well as higher emission intensities and higher emission permit prices. In the Intensity Target program, clean firms chose intensities significantly higher than predicted and produced significantly higher levels of output, leading to significantly higher emissions than expected while dirty firms did the opposite, choosing cleaner emission technologies, lower levels of output and emitting less than expected. In our experiment we enforced compliance by requiring participants to have permits on hand prior to production, and cash on hand in order to complete any transaction in the permit market. It would be interesting to compare the result when such provisions are not made, and instead participants are required to submit a report of their emissions. Cason and Gangadharan (2006), Murphy and Stranlund (2006), and Murphy and Stranlund (2007) explore compliance in emission permit markets under Cap and Trade programs. It may be the case that incentives under Intensity Targets are sufficiently different as to result in differing levels of compliance across programs and making one or the other better suited to emissions which cannot feasibly be perfectly monitored. The impact of relaxing the no-borrowing constraint imposed in our experiment may provide useful insights. Because under Intensity Targets dirty firms were limited by the number of permits created by clean firms it may be that enabling borrowing, either from firms' own future production or across firms in a futures market, will be sufficient to increase emissions to the

predicted levels and therefore undo any apparent gains attributable to Intensity Targets. The option of allowing for the purchase of shares entitling the owner to a stream of permits in each period as was done in the experiments by Muller and Mestelman (1994), Godby et al. (1997), Godby et al. (1998) and Mestelman et al. (1999) may alter the results, particularly in the Intensity Target environment.

Cap and Trade permit regulation was found to be more efficient than Intensity Target regulation, as predicted. However, several interesting findings accompany this result. First, if the regulator values emissions reduction above firm productivity, then for an equivalent emission target Intensity Targets would be preferred to the Cap and Trade because fewer emissions were produced. Second, although overall efficiency is higher under Cap and Trade, Intensity Targets yield higher output, so Intensity Targets maximize economic growth in the sector. Third, the only component of surplus responsible for the lower overall efficiency of Intensity targets is the producers surplus, which only affects the industry. If the industry advocates for the freedom of Intensity Targets rather than the strict quantity constraints of Cap and Trade, the regulator should support Intensity Targets. Fourthly, the calculation of efficiency assumes constant marginal damages. For a pollutant with marginal damages increasing in a sufficiently steep manner, Intensity Targets will be more efficient than Cap and Trade regulation. Finally, given that Intensity Targets do not require grandfathering or auctioning of permits and does not require banking, the results of this study suggest that regulators should consider the potential administrative costs avoided under Intensity Targets before opting for Cap and Trade.

Efficiency depends critically on the structure of the market for firms' output. This is because the output market is important in the calculation of total surplus, and thereby affects the efficiencies observed. Construction of a con-

trolled environment in which both the permit market and output markets are competitive would be a logical progression of work in this area. Such an environment would facilitate the consideration of an industry in which there may be clean and dirty firms independent of their emission intensity technologies and allow for the study of the evolution of an industry towards cleaner technologies when confronted with attempts to regulate emissions.

Banking decreased emission permit price variance and volatility in both Intensity Targets and Cap and Trade programs, significantly so for Intensity Targets, while increasing variance and volatility of aggregate emissions. This result is in line with those of Stranlund et al. (2014) and suggests that Cap and Trade without Banking is the most effective program for control of flow pollutants while stock pollutants could optionally be considered for control with Intensity Targets with Banking because for an equivalent level of expected emissions Intensity Targets achieve a similar level of efficiency and potentially encourage adoption of overall cleaner technologies.

References

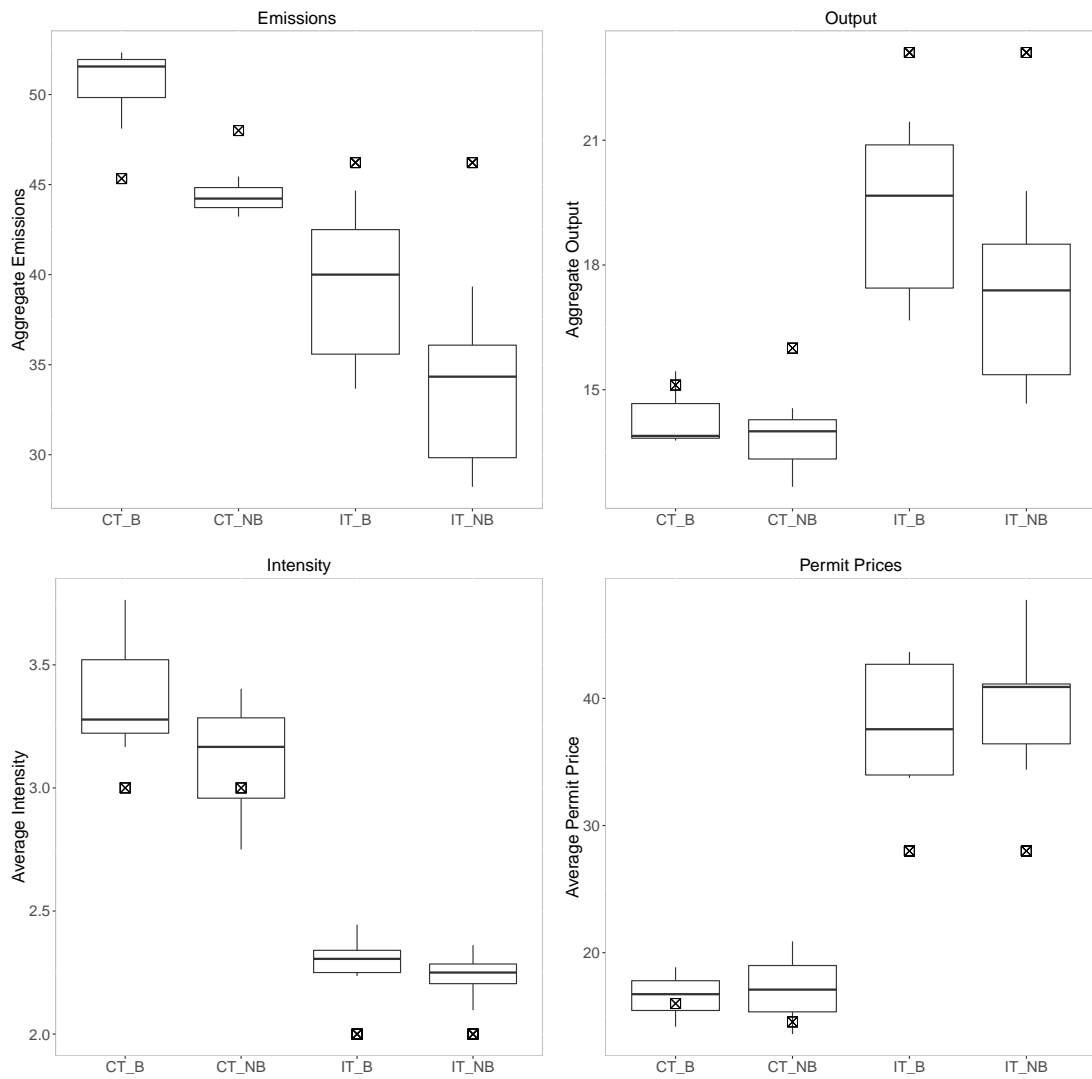
- Ben-David, S., Brookshire, D. S., Burness, S., McKee, M., and Schmidt, C. (1999). "Heterogeneity, Irreversible Production Choices, and Efficiency in Emission Permit Markets". In: *Journal of Environmental Economics and Management* 38.2, pp. 176–194.
- Ben-David, S., Brookshire, D., Burness, S., McKee, M., and Schmidt, C. (2000). "Attitudes Toward Risk and Compliance Emission Permit Markets." In: *Land Economics* 76.4, p. 590.
- Boom, J.-T. and Dijkstra, B. R. (2009). "Permit trading and credit trading: A comparison of cap-based and rate-based emissions trading under perfect and imperfect competition". In: *Environmental and Resource Economics* 44.1, pp. 107–136.
- Buckley, N. J., Mestelman, S., and Muller, R. A. (2006). "Implications Of Alternative Emission Trading Plans: Experimental Evidence". In: *Pacific Economic Review* 11.2, pp. 149–166.
- Buckley, N. J., Mestelman, S., and Muller, R. A. (2008). "Baseline-and-credit emission permit trading: experimental evidence under variable output capacity". In: *Environmental economics, experimental methods*. Ed. by T. L. Cherry, S. Kroll, and J. F. Shogren. Routledge.
- Buckley, N. J., Mestelman, S., and Muller, R. A. (2014). *Production Capacity and Abatement Technology Strategies in Emissions Trading Markets*. Manuscript 2014-16. McMaster.
- Cason, T. N. and Gangadharan, L. (2006). "Emissions variability in tradable permit markets with imperfect enforcement and banking". In: *Journal of Economic Behavior & Organization* 61.2, pp. 199–216.

- Cason, T. N. and Plott, C. R. (1996). "EPA's New Emissions Trading Mechanism: A Laboratory Evaluation". In: *Journal of Environmental Economics and Management* 30.2, pp. 133–160.
- Deweese, D. (2001). "Emissions Trading: ERCs or Allowances?" In: *Land Economics* 77.4, p. 513. ISSN: 00237639.
- Fell, H. and Morgenstern, R. D. (2010). "Alternative approaches to cost containment in a cap-and-trade system". In: *Environmental and Resource Economics* 47.2, pp. 275–297.
- Fell, H., Burtraw, D., Morgenstern, R. D., and Palmer, K. L. (2012). "Soft and hard price collars in a cap-and-trade system: A comparative analysis". In: *Journal of Environmental Economics and Management* 64.2, pp. 183–198. ISSN: 0095-0696.
- Fischbacher, U. (2007). "z-Tree: Zurich toolbox for ready-made economic experiments". In: *Experimental Economics* 10.2, pp. 171–178.
- Fischer, C. (2001). *Rebating environmental policy revenues: Output-based allocations and tradable performance standards*. Resources for the Future.
- Fischer, C. (2003). "Combining rate-based and cap-and-trade emissions policies". In: *Climate Policy* 3, S89–S103.
- Fischer, C. and Springborn, M. (2011). "Emissions targets and the real business cycle: Intensity targets versus caps or taxes". In: *Journal of Environmental Economics and Management* 62.3, pp. 352–366.
- Godby, R., Mestelman, S., Muller, R. A., and Welland, D. (1997). "Emissions Trading with Shares and Coupons when Control over Discharges Is Uncertain". In: *Journal of Environmental Economics and Management* 32.3, pp. 359–381.
- Godby, R., Mestelman, S., Muller, R. A., and Welland, D. (1998). "An experimental economic analysis of emissions trading with shares and coupons in the presence of market uncertainty". In: *Environmetrics* 9.1, pp. 67–79.

- Holt, C. and Laury, S. (2002). "Risk aversion and incentive effects". In: *American economic review* 92.5, pp. 1644–1655.
- Jotzo, F. and Pezzey, J. (2004). "Flexible targets for greenhouse gas emissions from developing countries under uncertainty". In: *Economics and Environment Network Working Paper, Australian National University, Canberra*.
- Jotzo, F. and Pezzey, J. (2007). "Optimal intensity targets for greenhouse gas emissions trading under uncertainty". In: *Environmental and Resource Economics* 38.2, pp. 259–284.
- Kirchkamp, O. (2013). *Utility to convert ztree to R*. URL: <http://www.kirchkamp.de/>.
- Marschinski, R. and Lecocq, F. (2006). *Do intensity targets control uncertainty better than quotas? Conditions, calibrations, and caveats*. Tech. rep. The World Bank.
- Mestelman, S., Moir, R., and Muller, R. A. (1999). "A Laboratory Test of Canadian Proposals for an Emission Trading Program". In: *Research in Experimental Economics*. Ed. by C. A. Holt and R. M. Isaac. Vol. 7. Greenwich, Connecticut: JAI Press, pp. 45–91.
- Muller, R. A. and Mestelman, S. (1994). "Emission Trading with Shares and Coupons: A Laboratory Experiment". In: *The Energy Journal* 15.2, pp. 185–211.
- Murphy, J. J. and Stranlund, J. K. (2006). "Direct and market effects of enforcing emissions trading programs: an experimental analysis". In: *Journal of Economic Behavior & Organization* 61.2, pp. 217–233.
- Murphy, J. J. and Stranlund, J. K. (2007). "A laboratory investigation of compliance behavior under tradable emissions rights: Implications for targeted enforcement". In: *Journal of Environmental Economics and Management* 53.2, pp. 196–212.

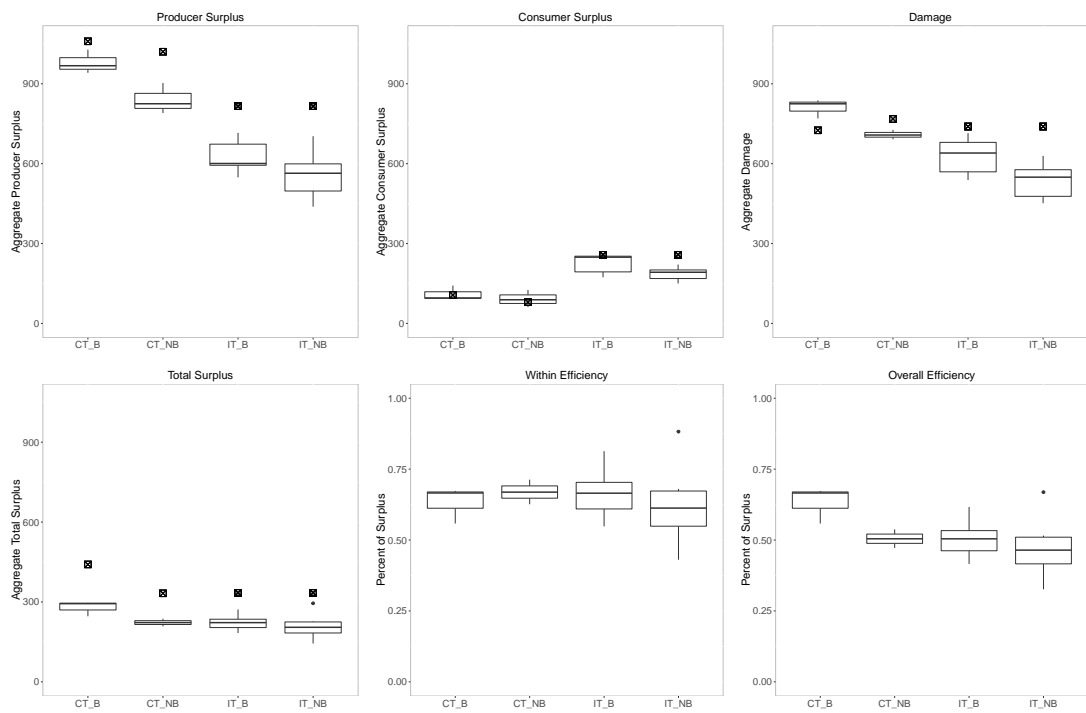
- OECD (2014). "The Cost of Air Pollution". In: DOI: <http://dx.doi.org/10.1787/9789264210448-en>. URL: </content/book/9789264210448-en>.
- Quirion, P. (2005). "Does uncertainty justify intensity emission caps?" In: *Resource and Energy Economics* 27.4, pp. 343–353.
- Schennach, S. M. (2000). "The Economics of Pollution Permit Banking in the Context of Title {IV} of the 1990 Clean Air Act Amendments". In: *Journal of Environmental Economics and Management* 40.3, pp. 189–210. ISSN: 0095-0696.
- Stranlund, J. K., Murphy, J. J., and Spraggon, J. M. (2011). "An experimental analysis of compliance in dynamic emissions markets". In: *Journal of Environmental Economics and Management* 62.3, pp. 414–429.
- Stranlund, J., Murphy, J., and Spraggon, J. (2014). "Price controls and banking in emissions trading: An experimental evaluation". In: *Journal of Environmental Economics and Management* 68.1, pp. 71–86.
- Sue Wing, I., Ellerman, A. D., and Song, J. (2006). *Absolute vs. intensity limits for CO2 emission control: performance under uncertainty*. Tech. rep. MIT Joint Program on the Science and Policy of Global Change.
- Tian, H. and Whalley, J. (2009). *Level versus equivalent intensity carbon mitigation commitments*. Tech. rep. National Bureau of Economic Research.
- Upadhyaya, P. (2010). "Is emission trading a possible policy option for India?" In: *Climate Policy* 10.5, pp. 560–574.

3.A Supplementary Figures



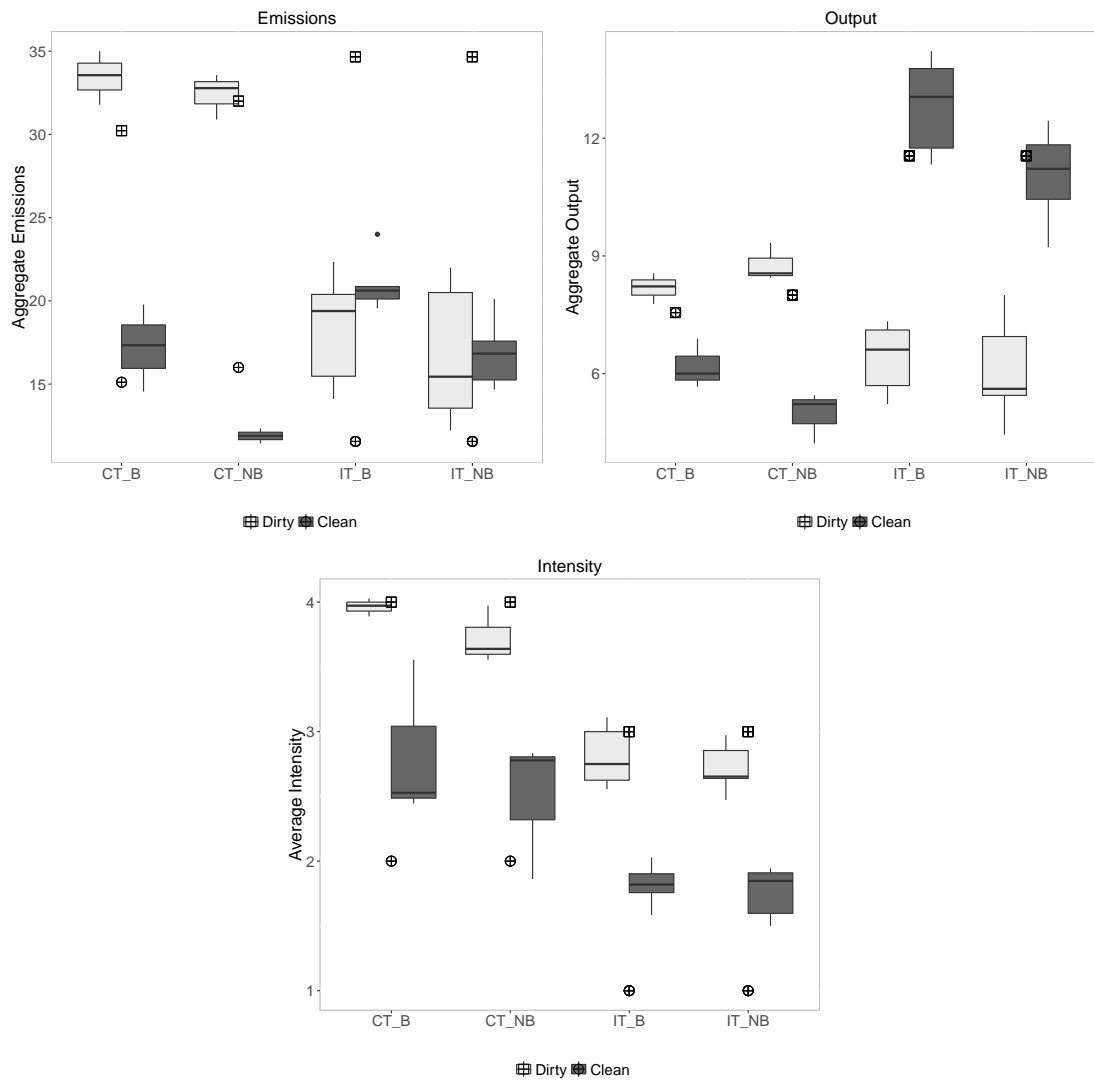
Boxes with crosses are predictions.

Figure 3.A.1: Box-plots of session results and predictions by treatment.



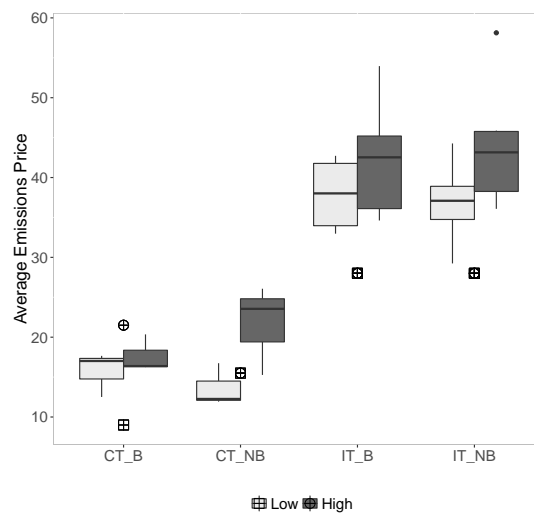
Boxes with crosses are predictions.

Figure 3.A.2: Box-plots of session efficiency results and predictions by treatment.



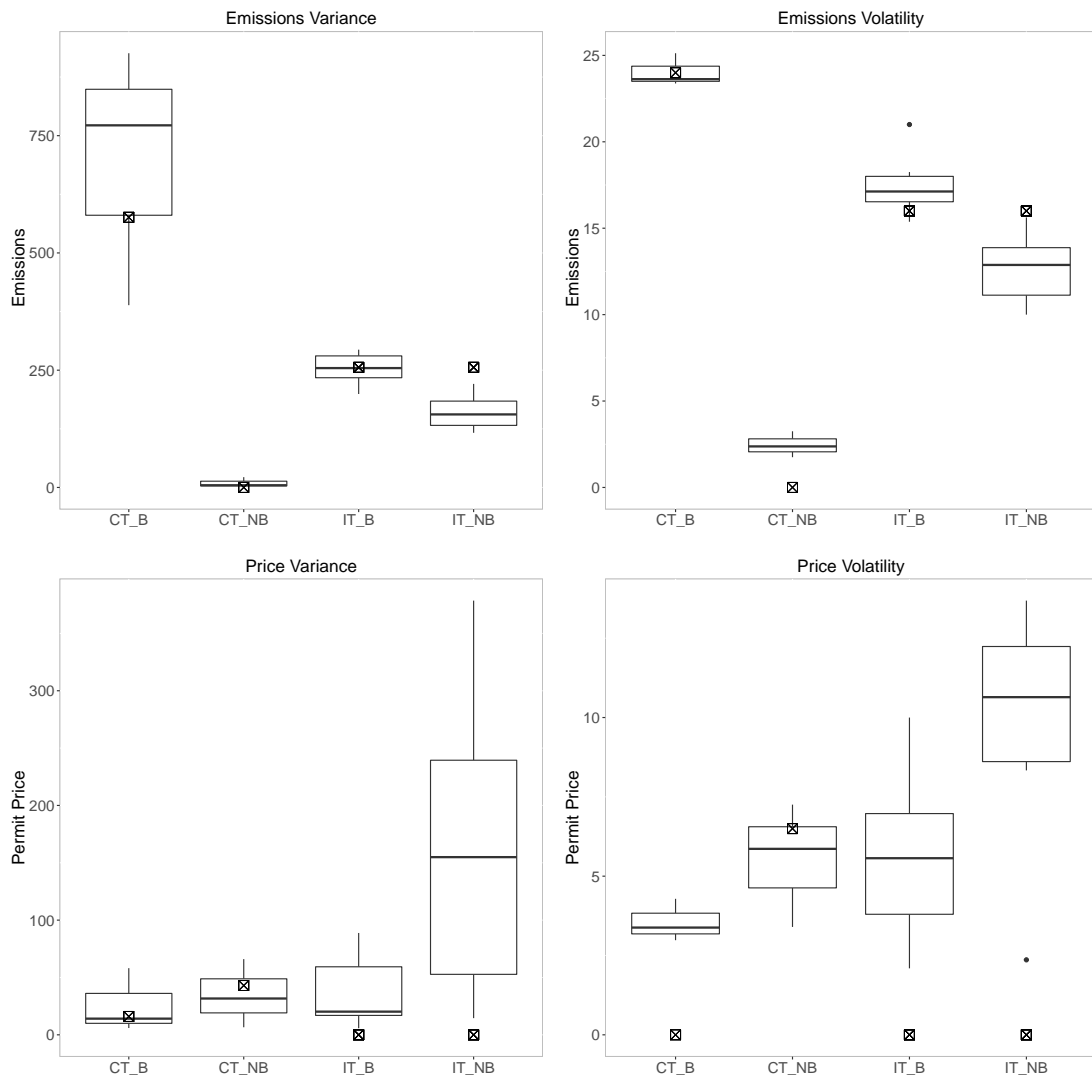
Boxes and circles with crosses are predictions.

Figure 3.A.3: Box-plots of session results and predictions by treatment and firm type.



Boxes and circles with crosses are predictions.

Figure 3.A.4: Box-plots of session emission permit price results and predictions by treatment and output demand state.



Boxes with crosses are predictions.

Figure 3.A.5: Box-plots of session volatility and variance results and predictions by treatment.

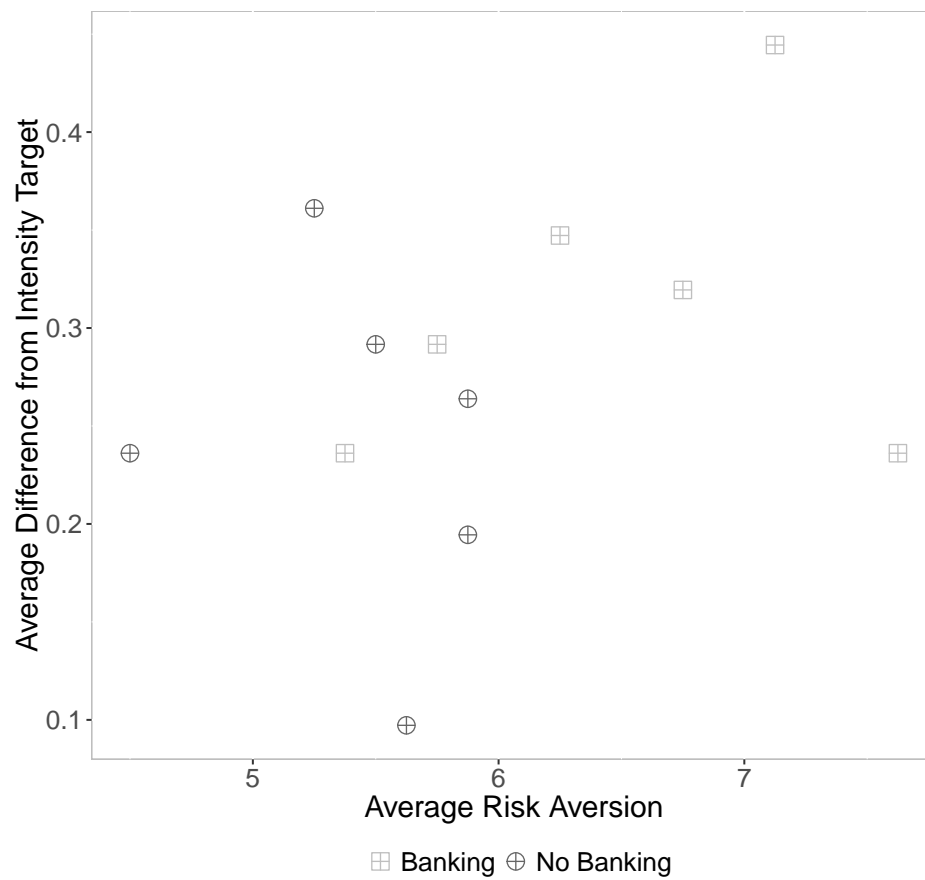


Figure 3.A.6: Average levels of risk aversion and deviation from the intensity target, by session.

3.B Experiment instructions and screen captures

Instructions IT_B

Stephanie Thomas

October 28, 2015

1 Introduction to Today's Session [TIMEI]

Welcome to the McMaster Experimental Economics Lab. In today's experiment you will have the opportunity to earn money that will be paid to you privately in cash at the end of today's session. Before we start the experiment we will conduct a short individual task.

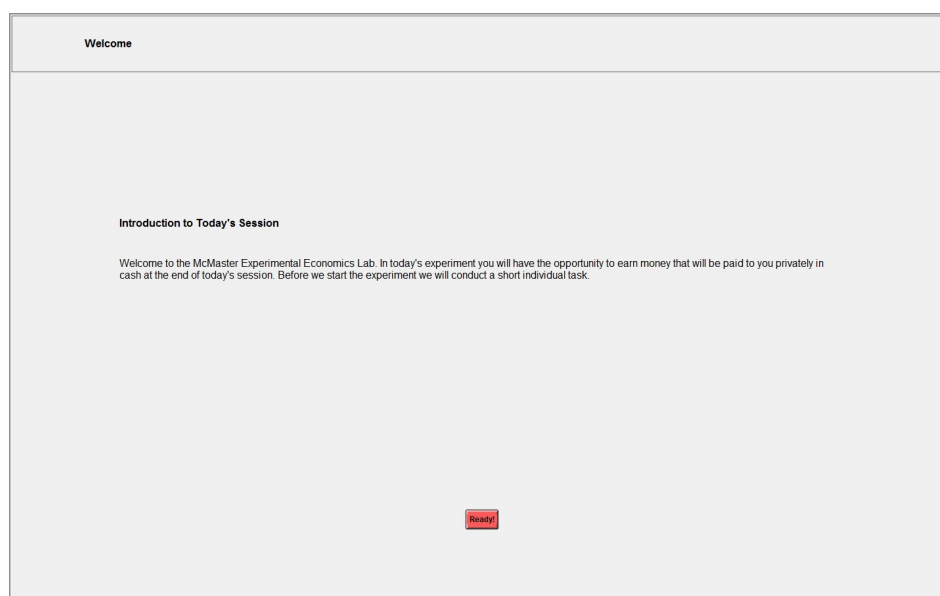


Figure 1: Introduction Screen

2 Individual Task [TIME1]

In this individual task we will ask you to make ten decisions. Each decision is a choice between 'Option A' and 'Option B'. One of the ten choices will be used in the end to determine your earnings for the survey. Before you start making your ten choices, please let me explain how these choices will affect your earnings for this part of the experiment.

After you make you make your 10 decisions between Option A and Option B using the computer software, the computer will pick two random numbers

between 1 and 10 (for example it might pick 1, 2, 3, 4, or so on up to 10). Each of the possibilities from 1 to 10 is equally likely so it will be like the computer is rolling two dice that each have 10 sides. The first random number will be used to select one of the ten decisions, which will be used to determine your survey payment. The second random number will be used to determine what your payoff is for the option you chose (A or B) for the particular decision selected. Even though you will make ten decisions, only one of these will end up affecting your task earnings and you will not know in advance which decision will be used. Each decision has an equal chance of being used in the end and these random numbers will be different for each participant in today's session.

Now, please look at the first decision at the top of your screen. Option A pays \$2.00 Canadian dollars if the random number is 1, and it pays \$1.60 if the number is between 2 and 10. Option B yields \$3.85 if the random number is 1, and it pays \$0.10 if the number is between 2 and 10. The other decisions are similar, except that as you move down the table on the screen, the chances of the higher payoff for each option increase. In fact, for Decision 10 in the bottom row, the random number will not be needed since each option pays the highest payoff for sure, so your choice here is between receiving \$2.00 in Option A or \$3.85 in Option B.

To summarize, you will make ten choices: for each decision row you will have to choose between Option A and Option B. You may choose A for some decision rows and B for other rows, and you may make your decisions in any order. When you are finished and click the 'DONE' button on the bottom of the screen, the computer will choose a random number between 1 and 10 to pick which of the 10 decisions to use as your payoff. Then the computer will choose a second random number between 1 and 10 to determine your money earnings for the Option you chose for that decision. You will not find out the results of the random numbers until the end of today's session after the experiment is completed. Your task earnings, in Canadian dollars, will be added to your experiment earnings and your show up fee, and you will be paid all earnings in cash when we finish the session.

Now please look at the empty circles in the centre of your computer screen. You will have to choose a decision, Option A or Option B by clicking each of circles. Now you may begin making your choices.

Please do not talk with anyone while we are doing this; raise your hand if you have a question. Once everyone is finished the task we will start the experiment.

Individual Task

A		Choice	B	
Probability 1, payoff 1	Probability 2, payoff 2		Probability 1, payoff 1	Probability 2, payoff 2
1/10, \$2.00	9/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	1/10, \$3.85	9/10, \$0.10
2/10, \$2.00	8/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	2/10, \$3.85	8/10, \$0.10
3/10, \$2.00	7/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	3/10, \$3.85	7/10, \$0.10
4/10, \$2.00	6/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	4/10, \$3.85	6/10, \$0.10
5/10, \$2.00	5/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	5/10, \$3.85	5/10, \$0.10
6/10, \$2.00	4/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	6/10, \$3.85	4/10, \$0.10
7/10, \$2.00	3/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	7/10, \$3.85	3/10, \$0.10
8/10, \$2.00	2/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	8/10, \$3.85	2/10, \$0.10
9/10, \$2.00	1/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	9/10, \$3.85	1/10, \$0.10
10/10, \$2.00	0/10, \$1.60	A <input type="radio"/> B <input type="radio"/>	10/10, \$3.85	0/10, \$0.10

Individual Task

In this individual task we will ask you to make ten decisions. Each decision is a choice between "Option A" and "Option B." One of the ten choices will be used in the end to determine your earnings for the task. Before you start making your ten choices, please let me explain how these choices will affect your earnings for this part of the experiment.

After you make your make your 10 decisions between Option A and Option B using the computer software, the computer will pick two random numbers between 1 and 10 (for example it might pick 1, 2, 3, 4, or so on up to 10). Each of the possibilities from 1 to 10 is equally likely so it will be like the computer is rolling two dice that each have 10 sides. The first random number will be used to select one of the ten decisions, which will be used to determine your payment in this task. The second random number will be used to determine what your payoff is for the option you chose (A or B) for the particular decision selected. Even though you will make ten decisions, only one of these will end up affecting your task earnings and you will not know in advance which decision will be used. Each decision has an equal chance of being used in the end and these random numbers will be different for each participant in today's session.

Now, please look at the first decision at the top of your screen. Option A pays \$2.00 Canadian dollars if the random number is 1, and it pays \$1.60 if the number is between 2 and 10. Option B yields \$3.85 if the random number is 1, and it pays \$0.10 if the number is between 2 and 10. The other decisions are similar, except that as you move down the table on the screen, the chances of the higher payoff for each option increase. In fact, for Decision 10 in the bottom row, the random number will not be needed since each option pays the highest payoff for sure, so your choice here is between receiving \$2.00 in Option A or \$3.85 in Option B.

To summarize, you will make ten choices: for each decision row you will have to choose between Option A and Option B. You may choose A for some decision rows and B for other rows, and you may make them in any order. When you are finished and click the "Done" button at the bottom of the screen, the computer will choose a random number between 1 and 10 to pick which of the 10 decisions to use as your payoff. Then the computer will choose a second random number between 1 and 10 to determine your money earnings for the Option you chose for that decision. You will not find out the results of the random numbers until the end of today's session after the experiment is completed. Your task earnings are in Canadian dollars and will be added to your experiment earnings and your show up fee. When we finish the experiment you will be paid all earnings in cash.

Now please look at the empty circles in the centre of your computer screen. You will have to choose a decision, Option A or Option B by clicking one of the circles in each pair. Now you may begin making your choices. Please do not talk with anyone while we are doing this, raise your hand if you have a question. Once everyone is finished the task we will start the experiment.

Figure 2: Individual Task Screen

3 Experiment Introduction [TIME2]

In today's experiment you have the opportunity to earn Lab dollars which will be converted to Canadian dollars at the end of the experiment, and paid to you privately in cash at the rate shown to you on your screen.

[3 sec pause]

Your decisions in the experiment, and those of the other 7 participants in the room today, will influence the amount of Lab dollars you earn. Before we begin the paid portion of the experiment you will first participate in six unpaid practice decision periods. The practice periods will give you a chance to learn about the experiment.

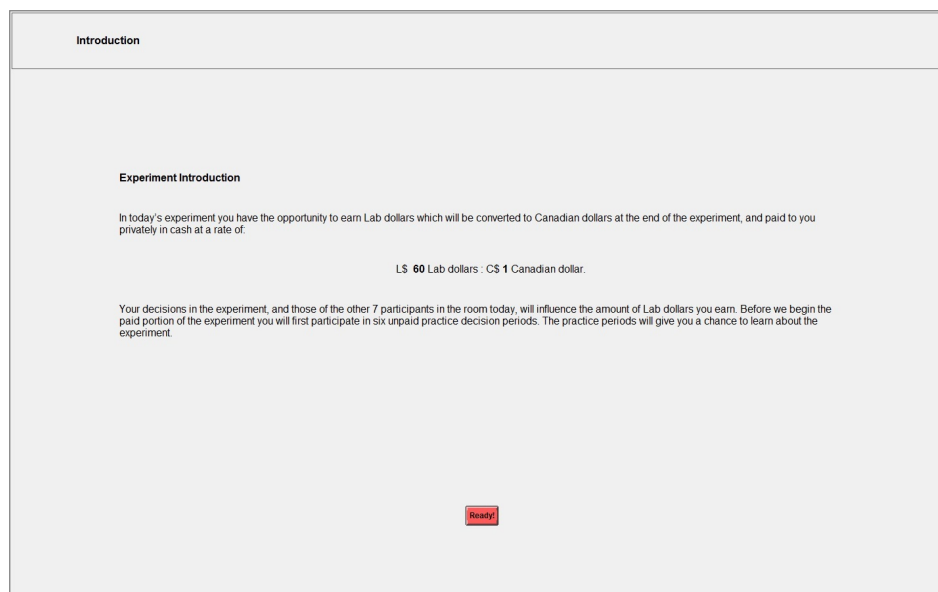


Figure 3: Experiment Introduction

4 Inputs and Outputs [TIME3]

In today's experiment you will act as a producer and you will turn Inputs into Outputs. Outputs are important because for each unit of Output you produce you will earn lab dollar revenues. Inputs are also important because in order to produce Output you will have to use Inputs. Each Output produced will require a specific number of Inputs to be used. We will call the ratio of the number of Inputs required to produce one unit of Output the 'Input Ratio', and we will always express this as:

units of Input : 1 unit of Output

Once you produce an Output, the revenues from that Output are added to your lab dollar balance and the Inputs required for production are removed from your inventory holdings right away.

Each participant will be given 3 free Inputs each time a unit of Output is produced (3 : 1 is your free input ratio). This means that if your Input Ratio is below 3 : 1, then you will gain Input each time you produce an Output. It is up to you to decide if you would like to increase your earnings by selling your Inputs to other producers or by buying Inputs and turning them into Outputs. Any unused Inputs you have not sold can be kept and used in the next Period.

An example might be useful. Suppose you have an Input Ratio of 4 : 1. This would mean that 4 Inputs are required to produce 1 unit of Output. Because you receive a 3 : 1 free Input Ratio, you will need to use 1 Input from your holdings to produce each unit Output. If you currently have 2 Inputs in your holdings then you can produce a maximum of 2 Outputs.

Suppose instead that your Input Ratio is 2 : 1. This would mean that 2 Inputs are required to produce 1 unit of Output. Because you receive the 3 : 1 Input Ratio for free, you would receive 1 input each time you produce a unit of Output. You can either sell this Input or you can keep it to use in the next period.

For the practice periods your Input Ratio will be either 2 : 1 or 4 : 1. Everyone will have the free Input Ratio of 3 : 1. Practice period earnings do not count towards your total cash earnings today so feel free to try things out.

Instructions

Inputs and Outputs

In today's experiment you will act as a producer and you will turn Inputs into Outputs. Outputs are important because for each unit of Output you produce you will earn lab dollar revenues. Inputs are also important because in order to produce Output you will have to use Inputs. Each Output produced will require a specific number of Inputs to be used. We will call the ratio of the number of Inputs required to produce one unit of Output the "Input Ratio", and we will always express this as:

units of **Input** : 1 unit of **Output**

Once you produce an Output, the revenues from that Output are added to your lab dollar balance and the Inputs required for production are removed from your inventory holdings right away.

Each participant will be given 3 free Inputs each time a unit of Output is produced (3:1 is your free input ratio). This means that if your Input Ratio is below 3:1, then you will gain Input each time you produce an Output. It is up to you to decide if you would like to increase your earnings by selling your Inputs to other producers or by buying Inputs and turning them into Outputs. Any unused Inputs you have not sold can be kept and used in the next decision period.

An example might be useful. Suppose you have an Input Ratio of 4:1. This would mean that 4 Inputs are required to produce 1 unit of Output. Because you receive a 3:1 free Input Ratio, you will need to use 1 Input from your holdings to produce each unit Output. If you currently have 2 Inputs in your holdings then you can produce a maximum of 2 Outputs.

Suppose instead that your Input Ratio is 2:1. This would mean that 2 Inputs are required to produce 1 unit of Output. Because you receive the 3:1 Input Ratio for free, you would receive 1 Input each time you produce a unit of Output. You can either sell this Input or you can keep it to use in the next period.

For the practice periods your Input Ratio will be either 2:1 or 4:1. Everyone will have the free Input Ratio of 3:1. Practice period earnings do not count towards your total cash earnings today so feel free to try things out.

When you are ready, click "Ready!" and wait for the slide show to begin.

Figure 4: Experiment Introduction: Inputs and Outputs

5 Introduction to the Market

5.1 jan_m_1 [20s]

This is an example of the 'Market' decision screen. This screen is used to buy Inputs from other participants, sell Inputs to other participants and also to produce Output. Let's begin by quickly introducing each section of the 'Market' decision screen and then we will go into detail afterward about how each section operates.

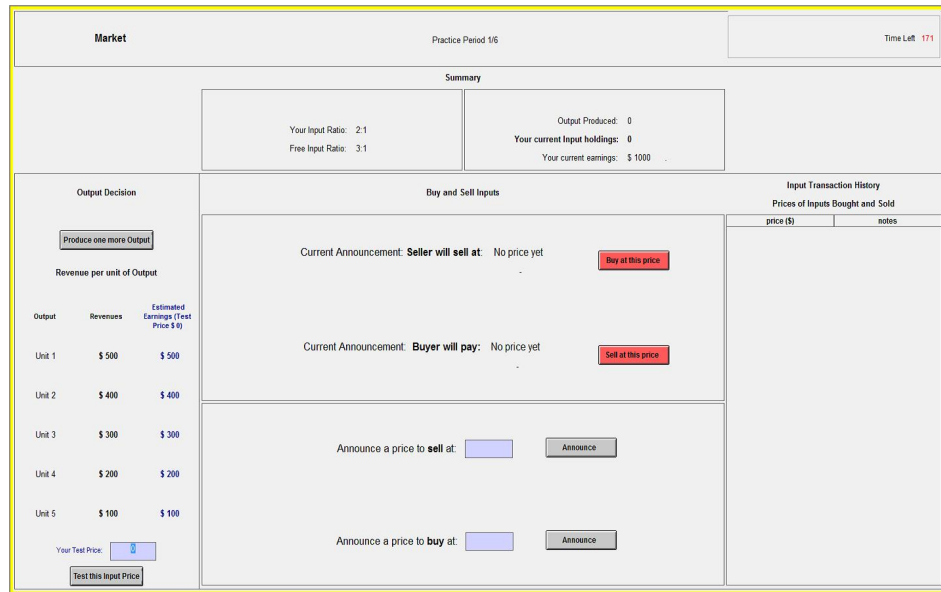


Figure 5: Market Introduction Screen 1

5.2 jan_m_2 [8s]

At the very top you will find the 'period information' section notifying you about the current period

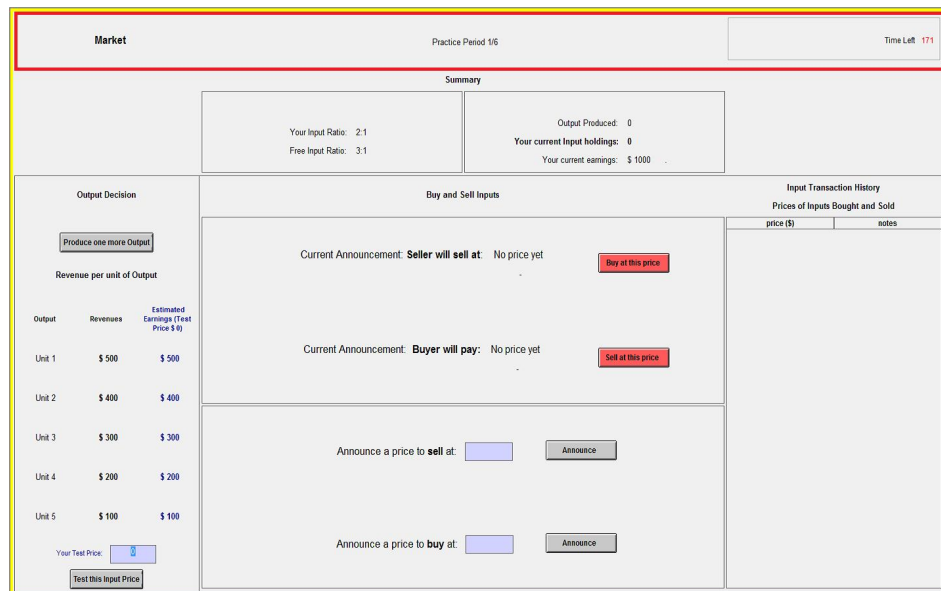


Figure 6: Market Introduction Screen 2

5.3 jan_m_3 [19]

Just beneath you will find the important 'summary section' which keeps track of important information such as your Input Ratio, the free Input Ratio, how

many Outputs you have produced, how many Inputs you currently own in your holdings and your current earnings.

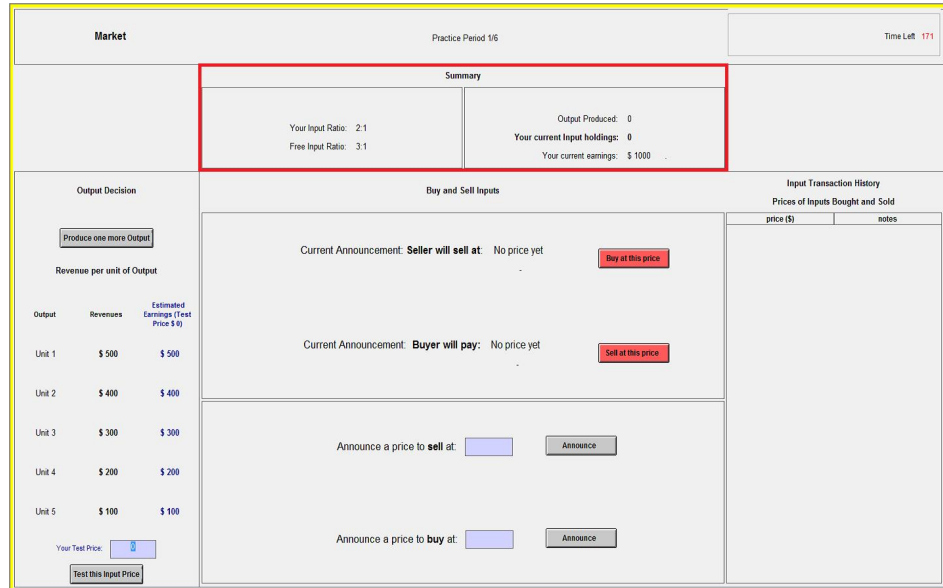


Figure 7: Market Introduction Screen 3

5.4 jan_m_4 [12]

At the bottom left of the 'Market' screen you will find the 'Output Decision' section, which displays revenue information and contains the button that you use to produce Output.

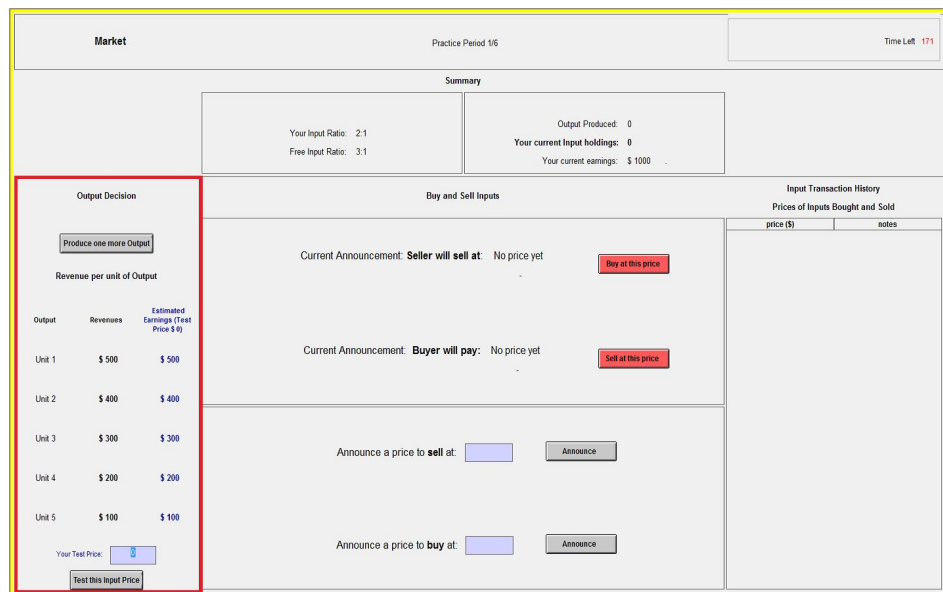


Figure 8: Market Introduction Screen 4

5.5 jan_m_5 [13]

At the bottom centre of the 'Market' screen you will find the 'Buy and Sell Inputs' section which allows you to buy and sell Inputs. We'll go into detail on how this section works later.

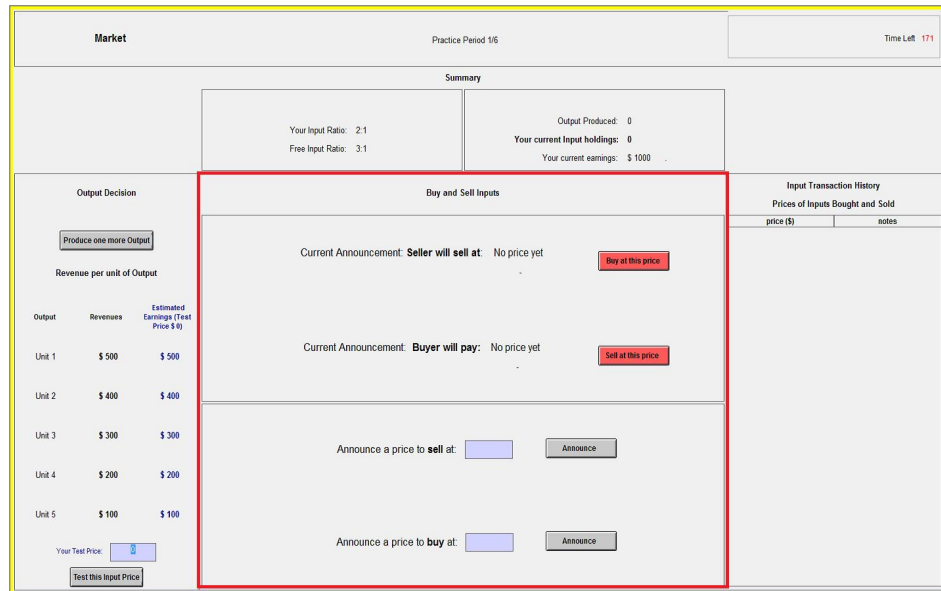


Figure 9: Market Introduction Screen 5

5.6 jan_m_6 [12s]

At the bottom right of the 'Market' screen you will find the 'Input Transaction History' section which will show you the prices of Inputs bought and sold during the Period.

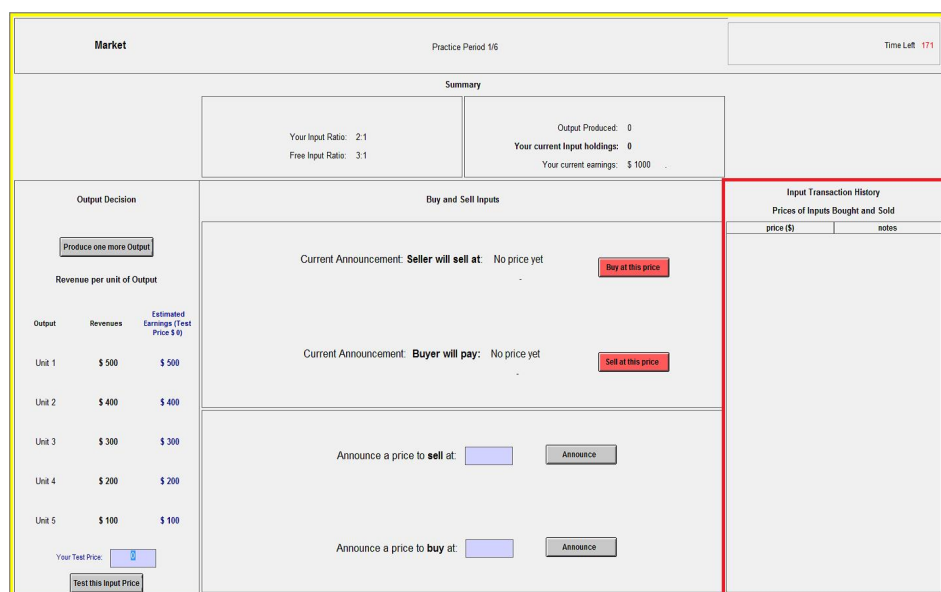


Figure 10: Market Introduction Screen 6

5.7 jan_m_7 [15s]

Now let's go back and discuss each section of the 'Market' decision screen in more detail. At the top you will find the Period number and the time remaining in this Period. In each practice period you will have 3 minutes to carry out your actions.

Market Practice Period 1/6 Time Left 171

Summary

Your Input Ratio: 2.1
Free Input Ratio: 3.1

Output Produced: 0
Your current Input holdings: 0
Your current earnings: \$ 1000

Output Decision

Produce one more Output

Revenue per unit of Output

Output	Revenues	Estimated Earnings (Test Price \$4)
Unit 1	\$ 500	\$ 500
Unit 2	\$ 400	\$ 400
Unit 3	\$ 300	\$ 300
Unit 4	\$ 200	\$ 200
Unit 5	\$ 100	\$ 100

Your Test Price: 0
Test this Input Price

Buy and Sell Inputs

Current Announcement: **Seller will sell at:** No price yet Buy at this price

Current Announcement: **Buyer will pay:** No price yet Sell at this price

Announce a price to sell at: [input field] Announce

Announce a price to buy at: [input field] Announce

Input Transaction History

Prices of Inputs Bought and Sold

price (\$)	notes

Figure 11: Market Introduction Screen 7

5.8 jan_m_8 [33s]

The Summary section displays 2 columns of important information. In the left column you will see your Input Ratio and your Free Input Ratio. This column will not change during the period. Notice that for the example currently shown on the screen it says that your Input Ratio is 2 : 1. In the next 3 practice periods some participants will have an Input Ratio of 2 : 1 and others will have an Input Ratio of 4 : 1. The 'Free Input Ratio' is the same for everyone, and will not change between rounds unless you are told otherwise.

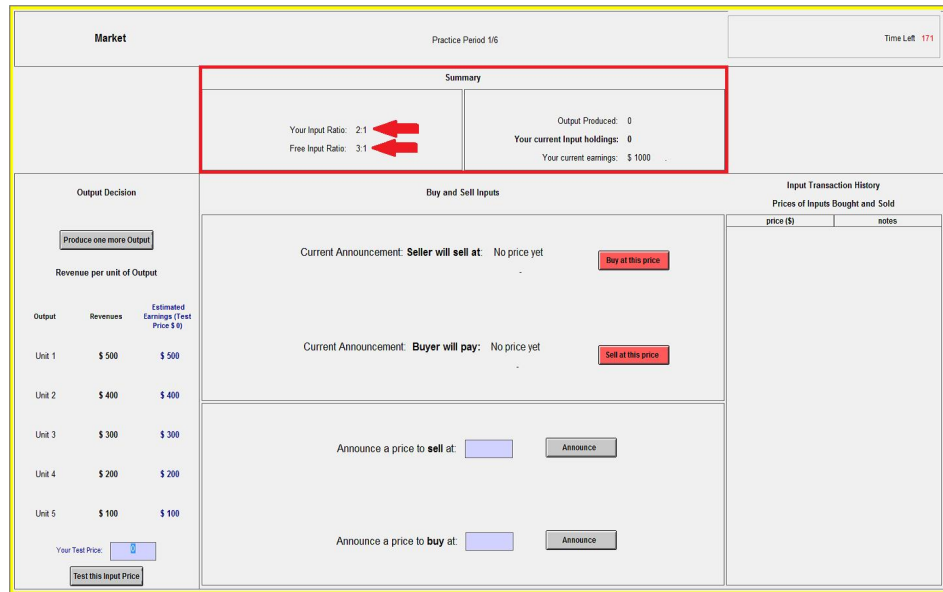


Figure 12: Market Introduction Screen 7

5.9 jan_m_9 [19]

The right column of the summary section shows how many Outputs you have produced this period, how many Inputs you currently have in your holdings and your current earnings. Everyone will begin the practice periods with L\$1000 Lab dollars and this will be reset to L\$250 just before we begin the first paid periods of today's experiment.



Figure 13: Market Introduction Screen 7

5.10 jan_m_10 [12]

In the section on the bottom left you will find the Output Decision section. For every unit of Output you produce you will earn the lab dollar Revenues shown to you in the table on the left.

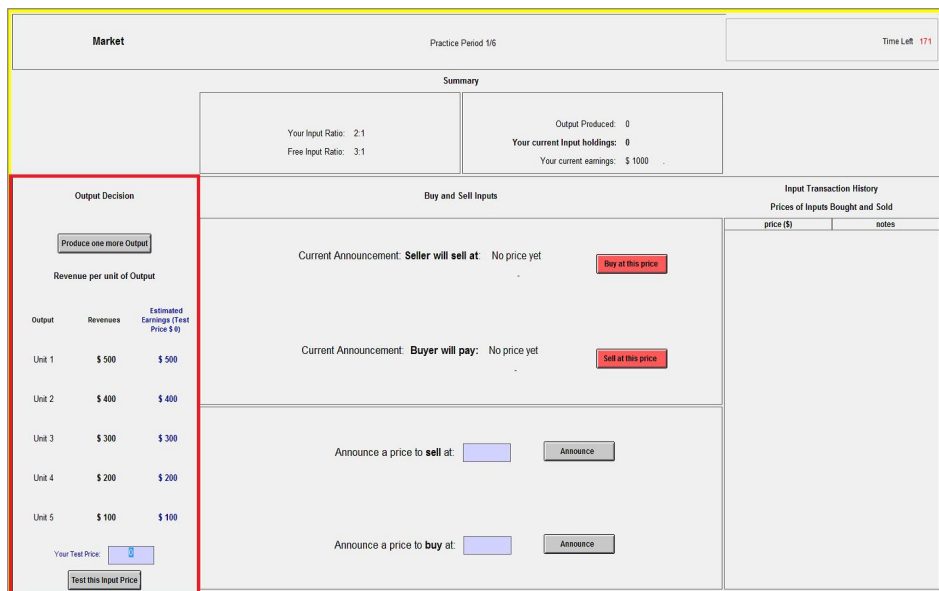


Figure 14: Market Introduction Screen 7

5.11 jan_m_11 [23]

You can see by looking down this table that the revenue you get for producing each additional unit of Output decreases as you produce more units of Output. This table will be used for all the Output you produce this period. After each period we reset this table back so you will again receive the revenue for Output unit 1 for the first Output you produce each period.

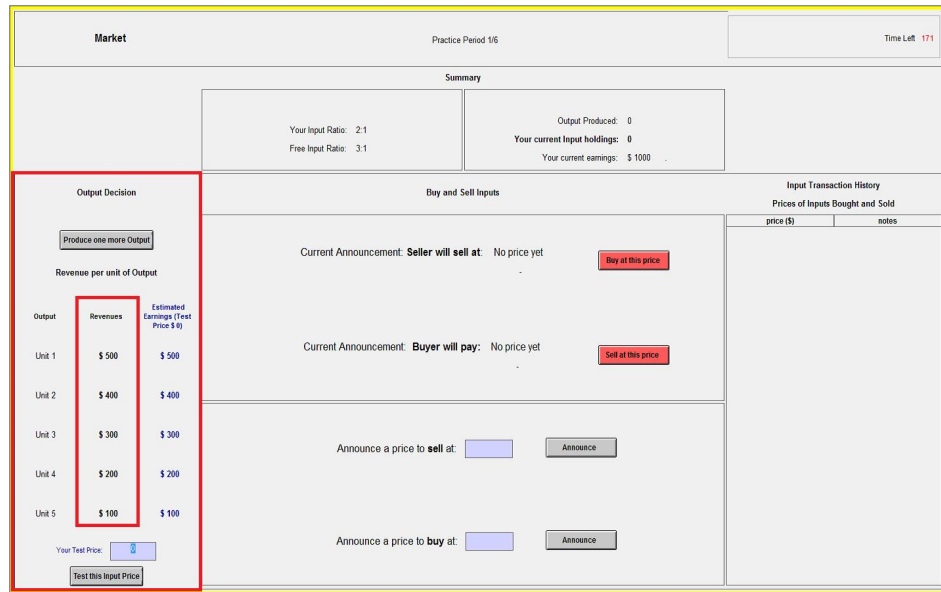


Figure 15: Market Introduction Screen 11

5.12 jan_m_12 [18]

If you want to produce a unit of Output click the ‘Produce one more Output’ at the top. Of course, you must have enough Inputs available to produce a unit of Output. You may need to buy more Inputs before producing Output. Let’s click the ‘Produce one more Output’ button to show you what happens.

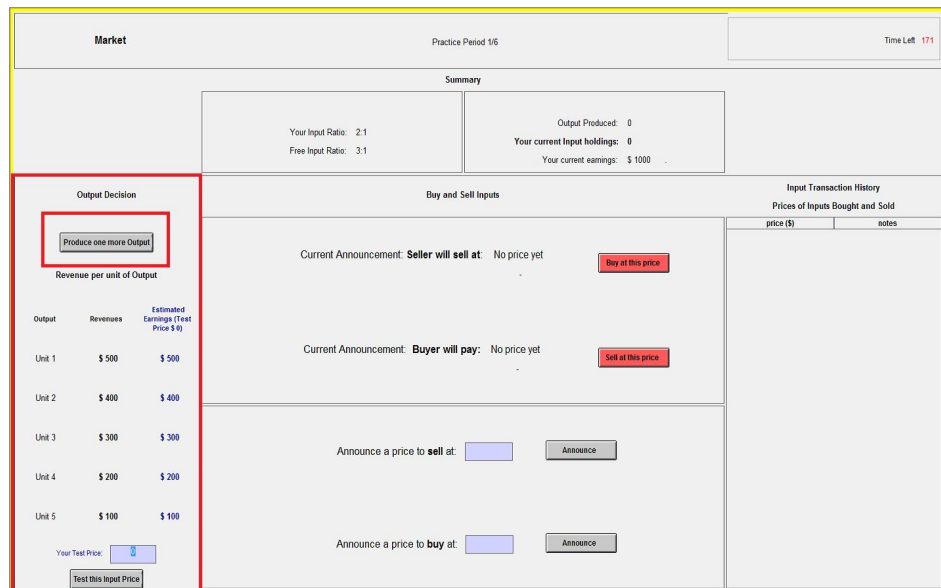


Figure 16: Market Introduction Screen 12

5.13 jan_m_13 [15]

Notice that as soon as you click the 'Produce one more Output' button the top row of the Output revenue table becomes bold signifying that you have produced your first unit of Output this period earning a revenue of L\$500.

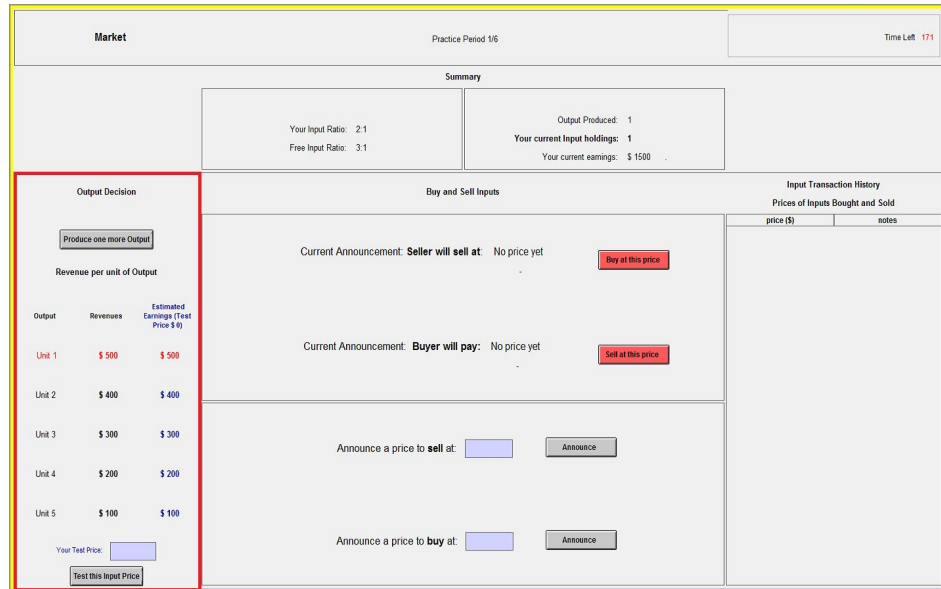


Figure 17: Market Introduction Screen 13.

5.14 jan_m_14 [12s]

In the Summary section your earnings have increased from your original L\$1000 by the L\$500 you just earned to give you a new current earnings of L\$1500.

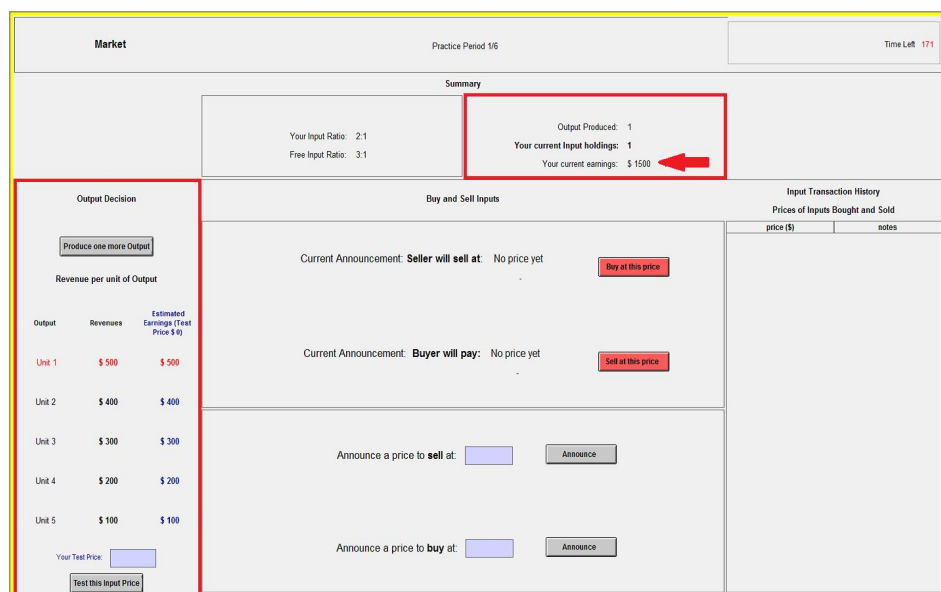


Figure 18: Market Introduction Screen 14

5.15 jan_m_15 [36s]

Also notice that you have received Input in accordance with your Input Ratio. Since your Input Ratio is 2 : 1, and the Free Input Ratio is 3 : 1, 'Your current Input Holdings' shown at the top of the screen has been updated to show you now have 1 Input in your holdings, up from the initial 0. If your Input Ratio were 4 : 1, you would need to have at least 1 Input in your Current Input holdings in order to produce the first unit of output. We will talk about how you can acquire more Inputs later. You could choose to sell the Input you have, or *you could keep it for use in a later Period.*

The screenshot shows the 'Market Introduction Screen 15' with the following details:

- Market Header:** Practice Period 1/6, Time Left: 171
- Summary:**
 - Your Input Ratio: 2:1
 - Free Input Ratio: 3:1
 - Output Produced: 1
 - Your current Input holdings: 1 (highlighted with a red arrow)
 - Your current earnings: \$ 1500
- Output Decision:**
 - Produce one more Output
 - Revenue per unit of Output

Output	Revenues	Estimated Earnings (Test Price \$1)
Unit 1	\$ 500	\$ 500
Unit 2	\$ 400	\$ 400
Unit 3	\$ 300	\$ 300
Unit 4	\$ 200	\$ 200
Unit 5	\$ 100	\$ 100

 - Your Test Price: [input field]
 - Test this Input Price
- Buy and Sell Inputs:**
 - Current Announcement: **Seller will sell at:** No price yet [Buy at this price]
 - Current Announcement: **Buyer will pay:** No price yet [Sell at this price]
 - Announce a price to sell at: [input field] [Announce]
 - Announce a price to buy at: [input field] [Announce]
- Input Transaction History:** Prices of Inputs Bought and Sold (price \$), notes

Figure 19: Market Introduction Screen 15

5.16 jan_m_16 [27s]

Notice that the 'Output Decision' section also contains an area in which you can enter price estimates of Inputs in order to test how they may affect the earnings you make on each unit of Output. Once you type an estimated Input price into the box you can click the 'Test this Input Price Estimate' to test the effects of estimated Input prices before you decide to produce a unit of Output, or buy or sell an Input.

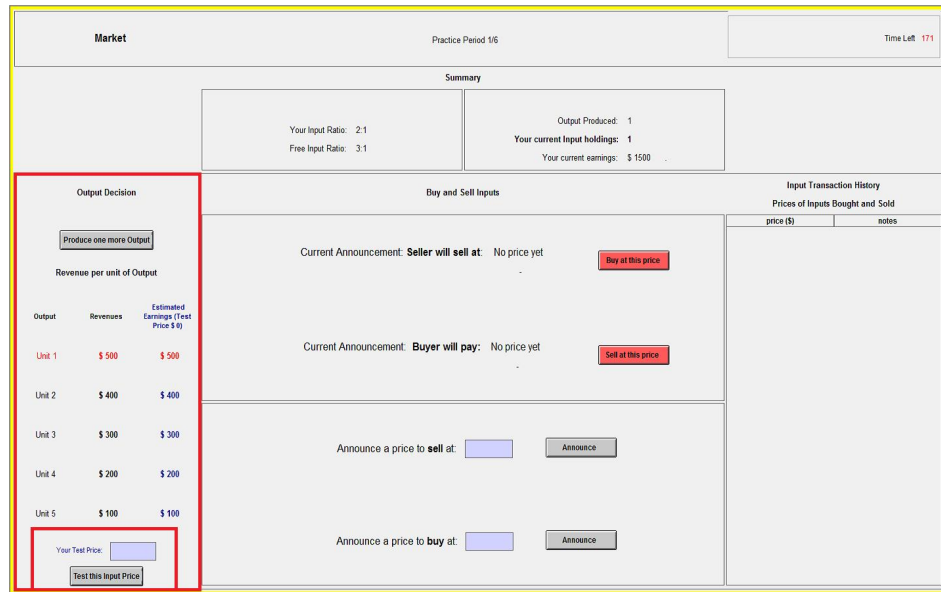


Figure 20: Market Introduction Screen 16

5.17 jan_m_17 [27w]

If your Input Ratio is above the Free Input Ratio, the blue numbers in the third column titled 'Estimated Earnings' are the result of subtracting the estimated Input costs per unit of Output from your Revenues. If your Input Ratio is below the Free Input Ratio, the blue numbers are the result of adding the estimated Input values per unit of Output, if you were to sell each input received at your Test Price.

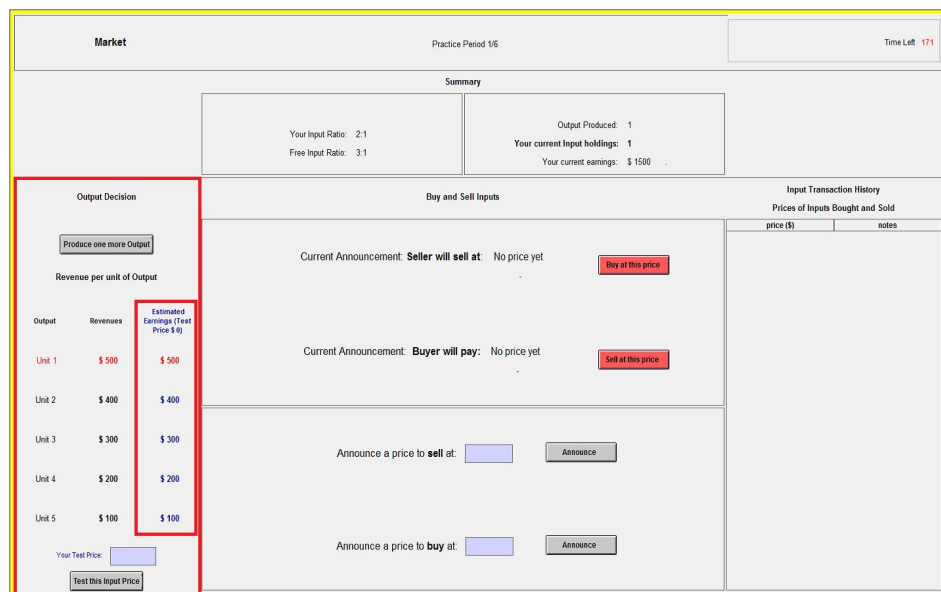


Figure 21: Market Introduction Screen 17

5.18 jan_m_18 [7]

For example, if you were to enter an Input price estimate of L\$150.00

Market Practice Period 1/6 Time Left 171

Summary

Your Input Ratio: 2:1
Free Input Ratio: 3:1

Output Produced: 1
Your current Input holdings: 1
Your current earnings: \$1500

Output Decision

Produce one more Output

Revenue per unit of Output

Output	Revenues	Estimated Earnings (Test Price \$4)
Unit 1	\$ 500	\$ 500
Unit 2	\$ 400	\$ 400
Unit 3	\$ 300	\$ 300
Unit 4	\$ 200	\$ 200
Unit 5	\$ 100	\$ 100

Your Test Price: 150
Test this Input Price

Buy and Sell Inputs

Current Announcement: **Seller will sell at:** No price yet

Current Announcement: **Buyer will pay:** No price yet

Announce a price to **sell** at:

Announce a price to **buy** at:

Input Transaction History

Prices of Inputs Bought and Sold	
price (\$)	notes

Figure 22: Market Introduction Screen 18

5.19 jan_m_19 [31]

The Test Centre will add the L\$150.00 estimated value for the one Input received to the revenue of each unit of Output produced (remember in this example your Input Ratio is 2 : 1 and the Free Input Ratio is 3 : 1). Looking at the first unit of Output you will see that with a revenue of L\$500 and one free Input with an estimated value of L\$150 your earnings would be L\$650. For your second Output your estimated earnings will be L\$550 if the price of each Input is \$150.

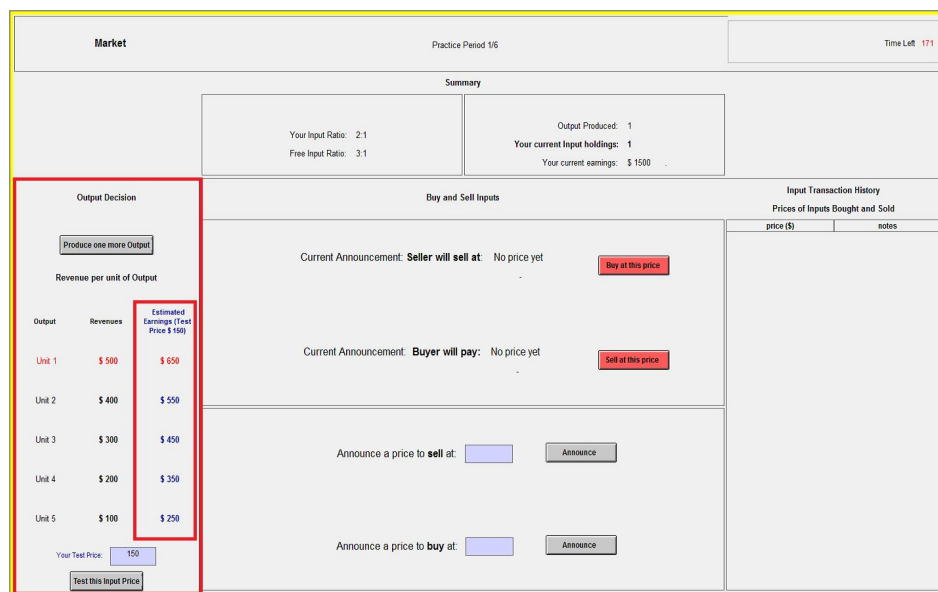


Figure 23: Market Introduction Screen 19

5.20 jan_m_20 [135]

Right now it is important to understand that you will be buying and selling Inputs in this experiment but you will not be buying and selling Outputs. Outputs are produced, which uses up Inputs and earns you revenue while Inputs can be bought in order to increase your Output production or Inputs can be sold to directly increase your earnings.

In the bottom middle section of the Market screen you will find the Buy and Sell Input section. Here you can both sell and buy Inputs to and from the other 7 participants in the today's session. Any Inputs that you have left over at the end of the period will be available to you in the following period, unless you are told otherwise.

You can buy and sell as many Inputs as you like during the 3 minute time allotted for each decision period but you can only buy and sell Inputs one at a time. Buying and selling Inputs is conducted in a similar way to trading on a stock market. Buyers announce prices they are willing to buy at and sellers announce prices they are willing to sell at. These offers are binding so that if another person wishes to sell an Input at a buyer's announced price then the seller can make the transaction. Similarly any person who wishes to buy an Input at a seller's announced selling price can also choose to make the transaction.

The market only allows for one buying price announcement and one selling announcement at a time. If someone wants to make a new buying announcement, they must announce a higher price than the current one. Similarly, if someone wants to make a new selling announcement, they must submit a lower price than the current one. To summarize, buying price announcements are only allowed to increase during a period and selling offers are only allowed to decrease during a period.

This is how announcements work, but when someone hears an announcement that appeals to them, they can buy or sell at the price announced by clicking on the appropriate button and a transaction will take place. This means that the buyer will gain an Input and will lose the agreed price from his or her earnings and the seller will lose an Input but will gain earnings equal to the agreed upon price.

As soon as the Input unit has been transacted the market will reset and the current buy and sell announcements will be deleted. Notice that the Buy and Sell Inputs section at the bottom of the screen is divided into 2 sections.

Market Practice Period 1/6 Time Left 171

Summary

Your Input Ratio: 2:1
Free Input Ratio: 3:1

Output Produced: 1
Your current Input holdings: 1
Your current earnings: \$ 1500

Output Decision

Produce one more Output

Revenue per unit of Output

Output	Revenues	Estimated Earnings (Net Price \$150)
Unit 1	\$ 500	\$ 650
Unit 2	\$ 400	\$ 550
Unit 3	\$ 300	\$ 450
Unit 4	\$ 200	\$ 350
Unit 5	\$ 100	\$ 250

Your Test Price: 150
Test this Input Price

Buy and Sell Inputs

Current Announcement: **Seller will sell at:** No price yet

Current Announcement: **Buyer will pay:** No price yet

Announce a price to **sell** at:

Announce a price to **buy** at:

Input Transaction History
Prices of Inputs Bought and Sold

price (\$)	notes
------------	-------

Figure 24: Market Introduction Screen 20

5.21 jan_m_21 [4]

The bottom section is devoted to making price announcements. [Pause]

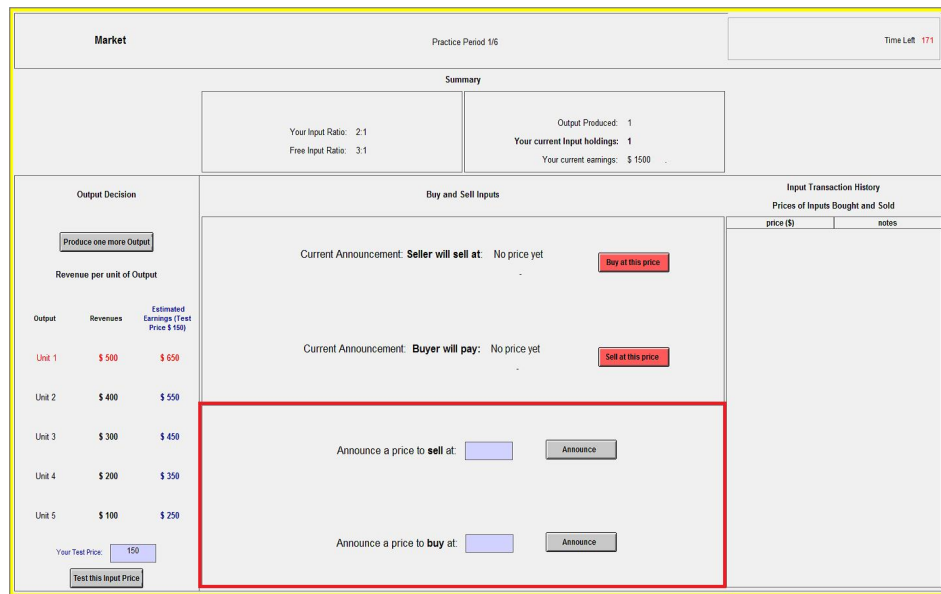


Figure 25: Market Introduction Screen 21

5.22 jan_m_22 [7s]

... and the top section is devoted buying and selling Inputs at the announced price. [Pause]

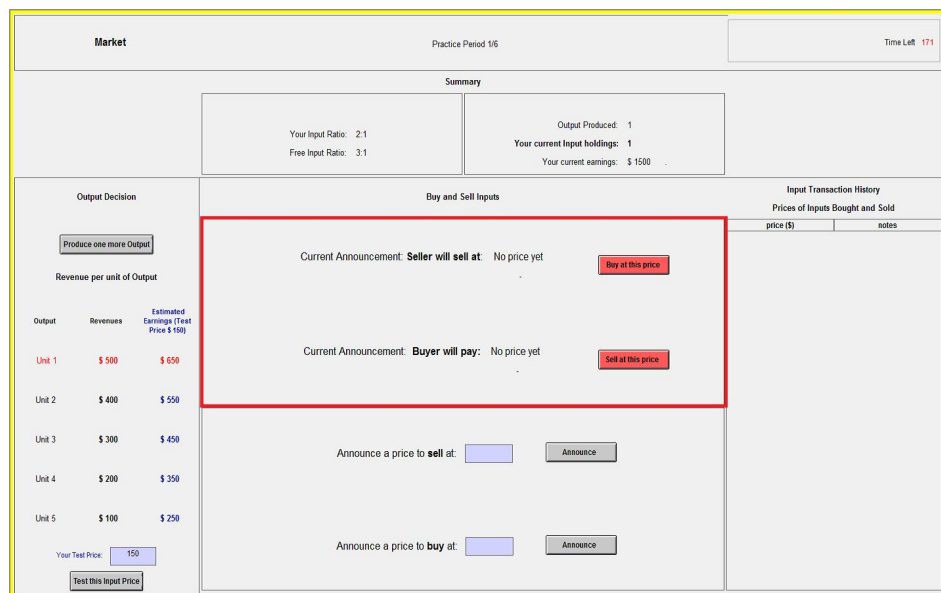


Figure 26: Market Introduction Screen 22

5.23 jan_m_23 [19s]

Going back to our example, suppose you want to try to sell your Input so that you can increase your earnings. Let's assume that you are willing to sell for less than the current seller's announcement of L\$380 but you would like to see

if you can get more than the current buyer’s announcement of L\$300. Suppose that you would like to announce a price of L\$345.

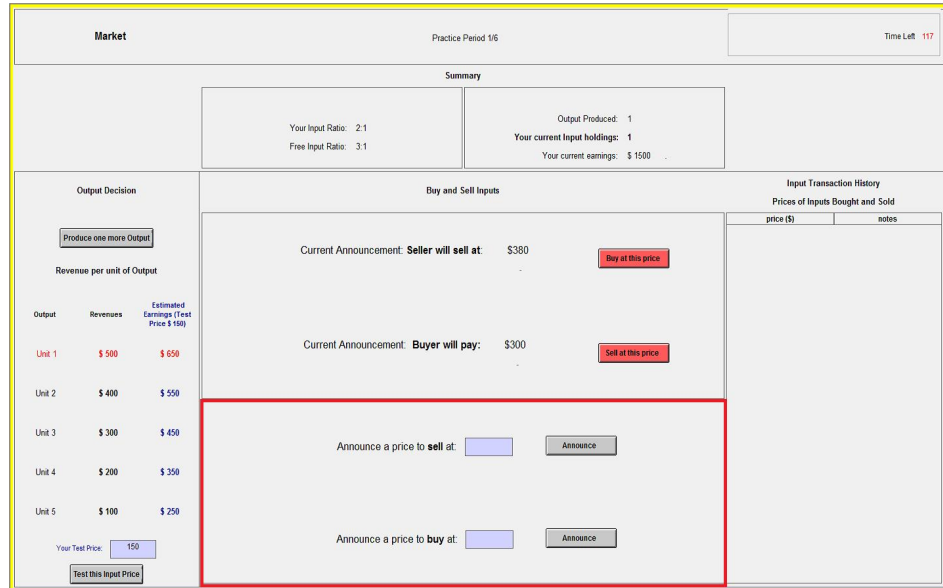


Figure 27: Market Introduction Screen 23

5.24 jan_m_24 [5s]

To do this enter 345 in the upper box of the lower section.

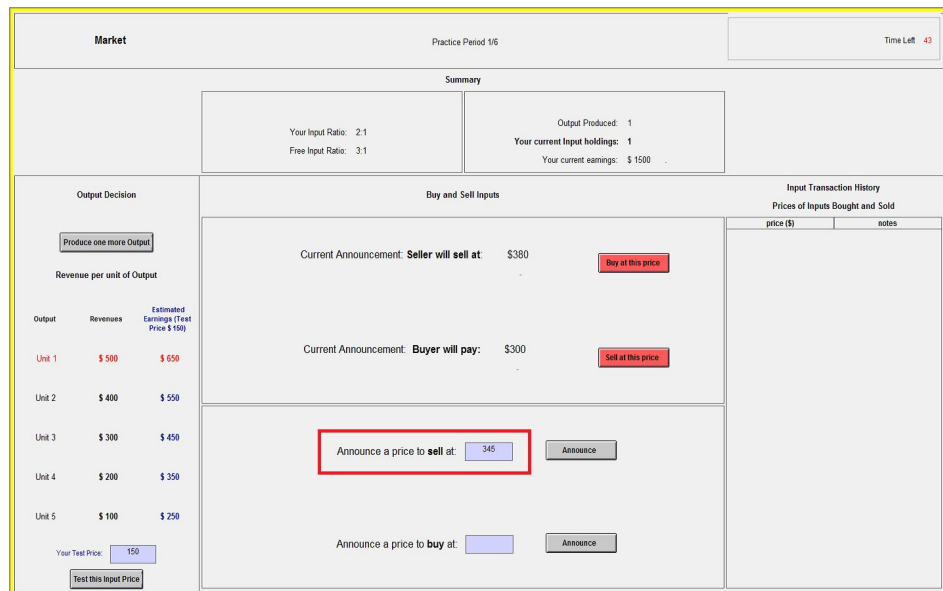


Figure 28: Market Introduction Screen 24

5.25 jan_m_25 [3s]

and then click the 'Announce' button.

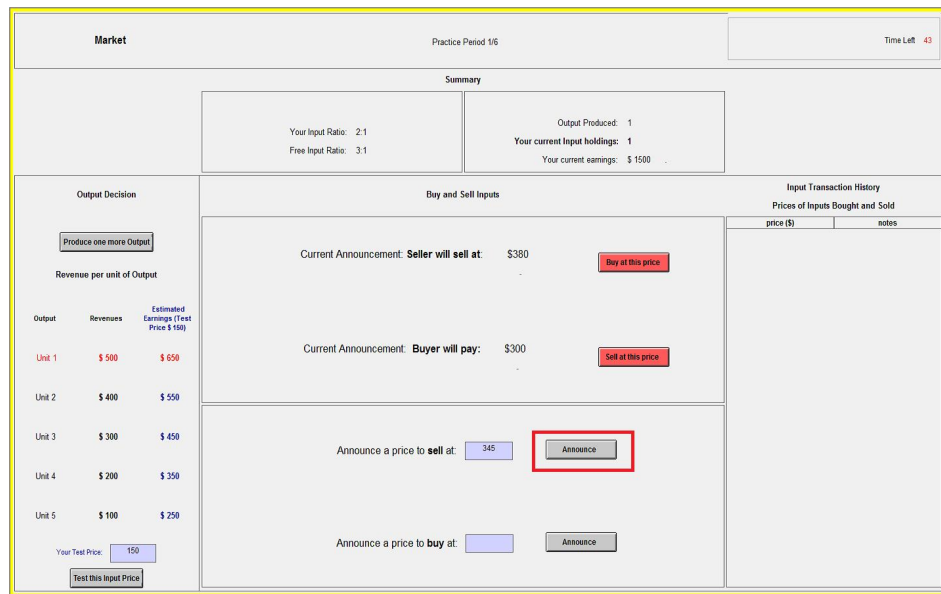


Figure 29: Market Introduction Screen 25

5.26 jan_m_26 [40s]

Notice how the current selling price announcement shown in the top has now changed from L\$380 to L\$345. Also notice that there is now a label below the L\$345 telling you that this is your announcement (All transactions will remain anonymous, so other participants will not see this message, and you will not see these messages of other participants.). When you announce a new selling price you must have at least 1 Input in your current holdings to sell. Simply making an announcement does not guarantee you will sell an Input. You will only sell if another participant is willing to buy from you at your announced price. Also, you cannot 'take-back' an announcement once it has been made even if you change your mind.

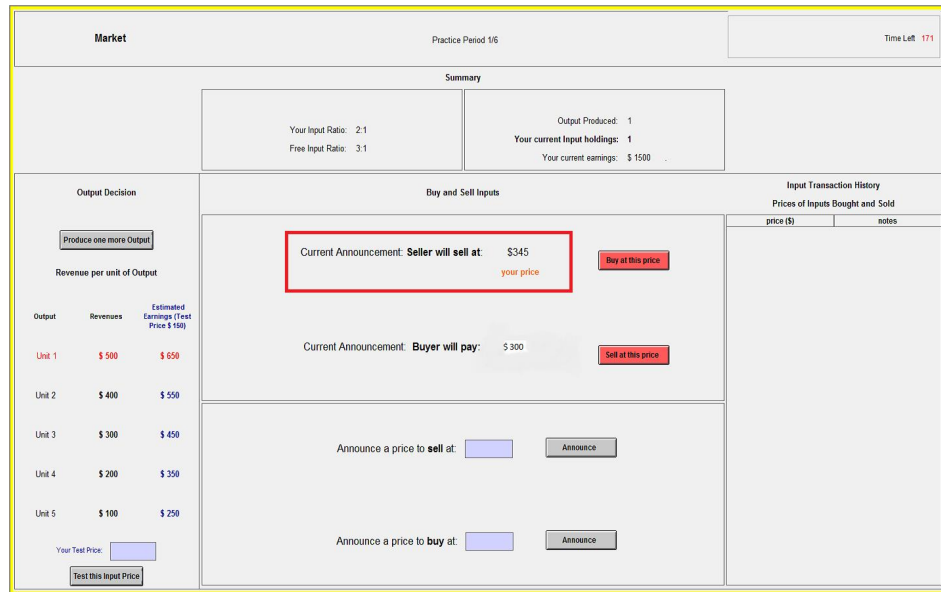


Figure 30: Market Introduction Screen 26

5.27 jan_m_27 [17s]

Now let's assume that you really wanted to sell that Input but no one was willing to buy from you for L\$345. Suppose you were willing to sell at the current buyer's announced price of L\$300 that is displayed in the upper section. Whenever you want to sell an Input at the currently announced price you

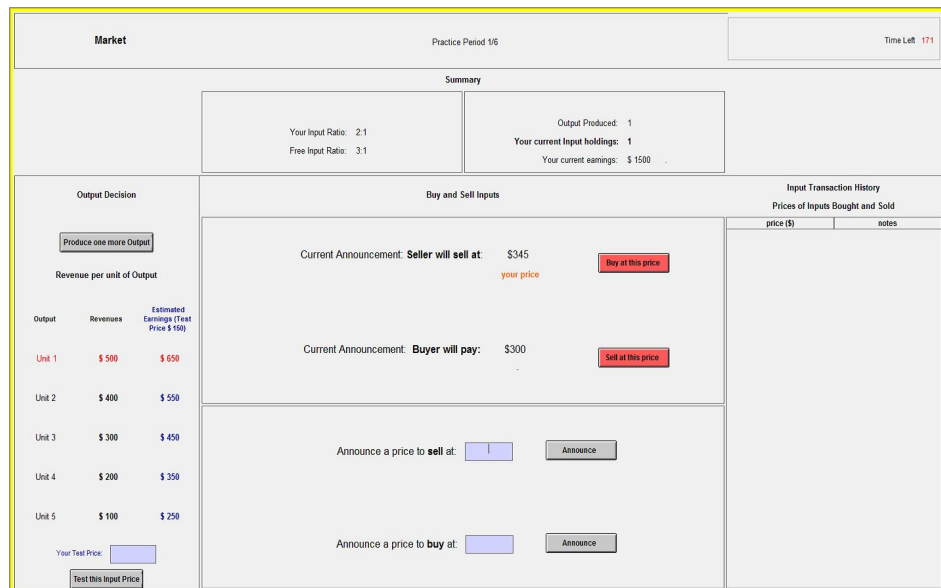


Figure 31: Market Introduction Screen 27

5.28 jan_m_28 [5s]

just click the 'Sell at this price' button on the top.

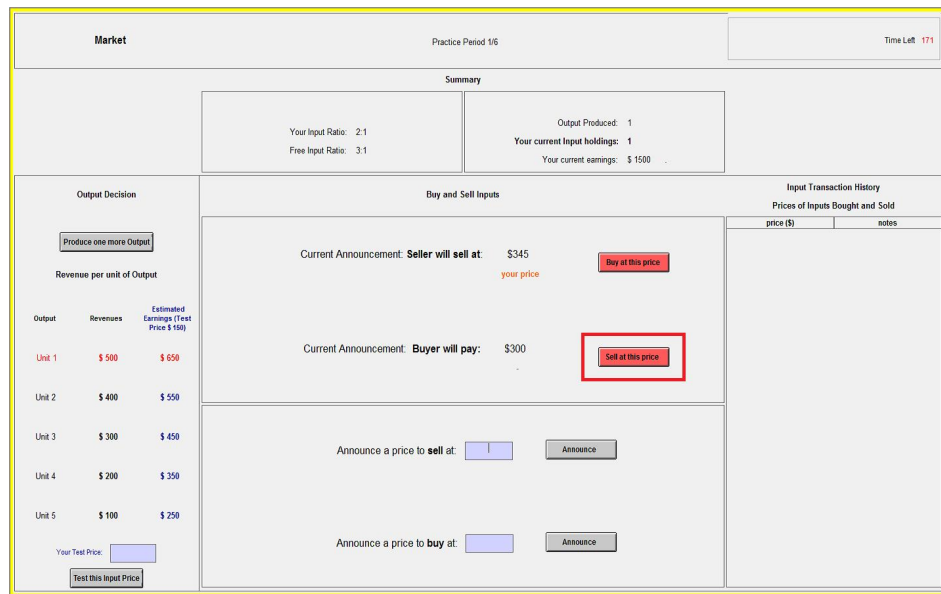


Figure 32: Market Introduction Screen 28

5.29 jan_m_29 [26s]

Notice that when we do this the transaction instantly takes place between you and the buyer who made the current announcement. The summary section at the top of the screen tells you that you now have 0 Inputs in your current holdings instead of 1 and that your current earnings have gone up by the \$300 sale price from L\$1500 to L\$1800. Also notice that the current buying and selling announcements have been reset and the market is ready for buying and selling the next Input.

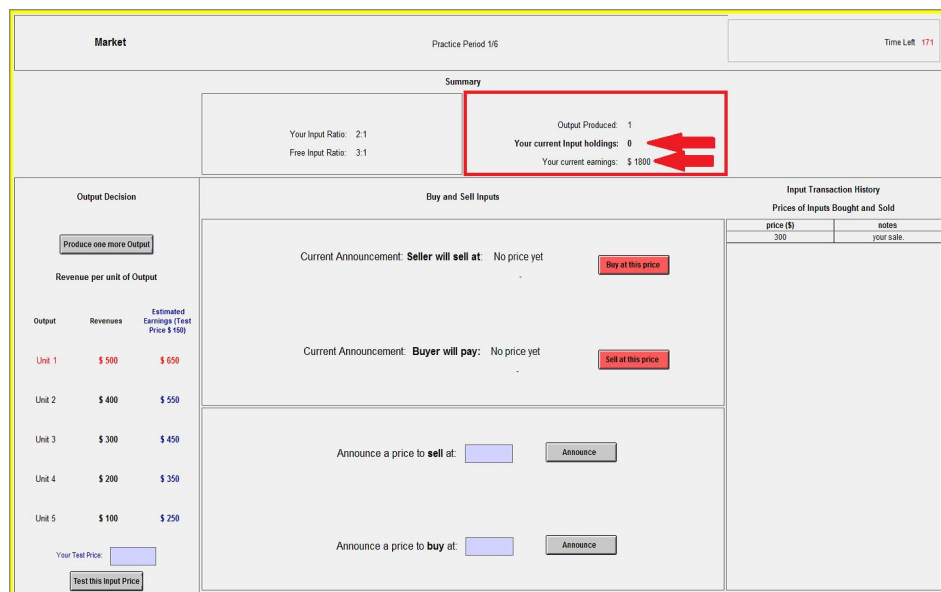


Figure 33: Market Introduction Screen 29

5.30 jan_m_30 [30s]

In addition, whenever an Input has been bought or sold, the price of the transaction will appear on each participant's 'Input Transaction History' at the bottom right of the screen. When you buy a unit you will see a 'your buy' note, and when you sell, you will see the 'your sale' note in the rightmost column. This is for your information only, no one else will know who bought or sold as the trading identities of others in the experiment are kept anonymous. Although this example shows how to sell an Input, you can also buy an Input in much the same way.

Market Practice Period 1/6 Time Left 171

Summary

Your Input Ratio: 2:1
Free Input Ratio: 3:1

Output Produced: 1
Your current Input holdings: 0
Your current earnings: \$1800

Output Decision

Produce one more Output

Revenue per unit of Output

Output	Revenues	Estimated Earnings (Test Price \$150)
Unit 1	\$ 500	\$ 650
Unit 2	\$ 400	\$ 550
Unit 3	\$ 300	\$ 450
Unit 4	\$ 200	\$ 350
Unit 5	\$ 100	\$ 250

Your Test Price:
Test this Input Price

Buy and Sell Inputs

Current Announcement: **Seller will sell at:** No price yet

Current Announcement: **Buyer will pay:** No price yet

Announce a price to **sell** at:

Announce a price to **buy** at:

Input Transaction History

Prices of Inputs Bought and Sold

price (\$)	notes
300	your sale

Figure 34: Market Introduction Screen 30

5.31 jan_m_31 [20]

Forgetting about the previous example, suppose that you really want to buy 1 Input (like you would if your Input Ratio was 4 : 1 and you wanted to produce one unit of Output). There are currently no buy or sell announcements so you can choose whatever price you like. Suppose you wanted to announce you were willing to buy the Input for L\$250.

Figure 35: Market Introduction Screen 31

5.32 jan_m_32 [6s]

To do this enter 250 in the bottom box of the lower section and click the 'Announce' button.

Figure 36: Market Introduction Screen 32

5.33 jan_m_33 [39s]

Notice how the current buying price announcement shown in the top has now changed from nothing to L\$250. Also notice that there is now a label below the L\$250 telling you that this is your announcement. When you announce a new

buying price you must have enough current earnings to be able to buy the unit at that price. Simply making a buying announcement does not guarantee you will buy an Input. You will only buy if another participant is willing to sell to you at your announced price. Also, you cannot 'take-back' an announcement once it has been made even if you change your mind. Any price you announce to buy at must be higher than the current buyer's announced price displayed in the top section.

Market Practice Period 1/6 Time Left 171

Summary

Your Input Ratio: 2:1
Free Input Ratio: 3:1

Output Produced: 1
Your current Input holdings: 0
Your current earnings: \$ 1800

Output Decision

Produce one more Output

Revenue per unit of Output

Output	Revenues	Estimated Earnings (Last Price \$ 500)
Unit 1	\$ 500	\$ 650
Unit 2	\$ 400	\$ 550
Unit 3	\$ 300	\$ 450
Unit 4	\$ 200	\$ 350
Unit 5	\$ 100	\$ 250

Your Test Price:
Test this Input Price

Buy and Sell Inputs

Current Announcement: **Seller will sell at:** No price yet

Current Announcement: **Buyer will pay:** \$250
your price

Announce a price to sell at:

Announce a price to buy at:

Input Transaction History

Prices of Inputs Bought and Sold	
price (\$)	notes
300	your sale

Figure 37: Market Introduction Screen 33

5.34 jan_m_34 [21s]

Although there are no current selling announcements, let's assume that another subject suddenly made an announcement he or she was willing to sell at a price of L\$275. Furthermore, let's assume after a while you don't think anyone will sell at your announced price of L\$250 before the end of the period and you decide you would like to buy at the current price of L\$275.

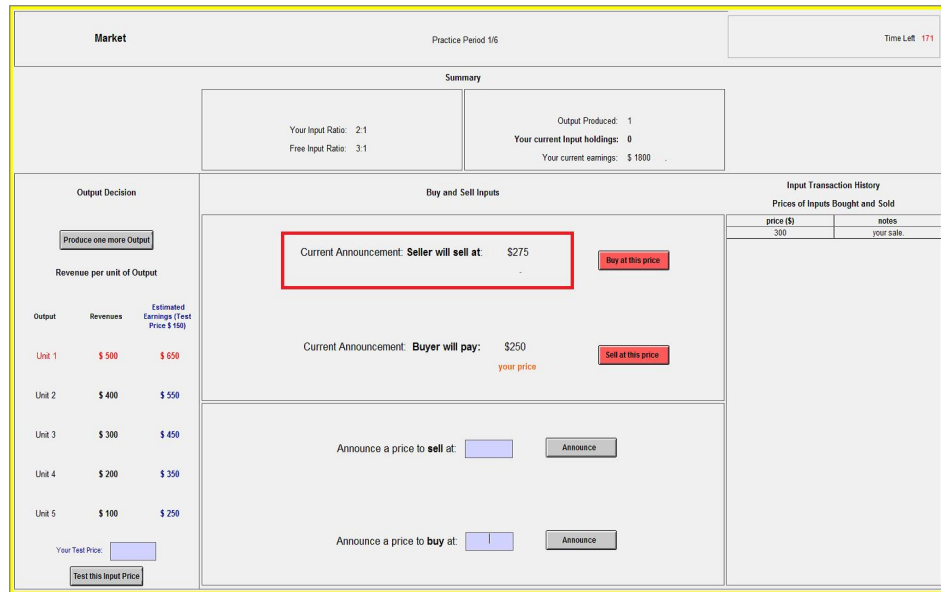


Figure 38: Market Introduction Screen 34

5.35 jan_m_35 [5s]

To do this you simply click the 'Buy at this price' button on the top section.

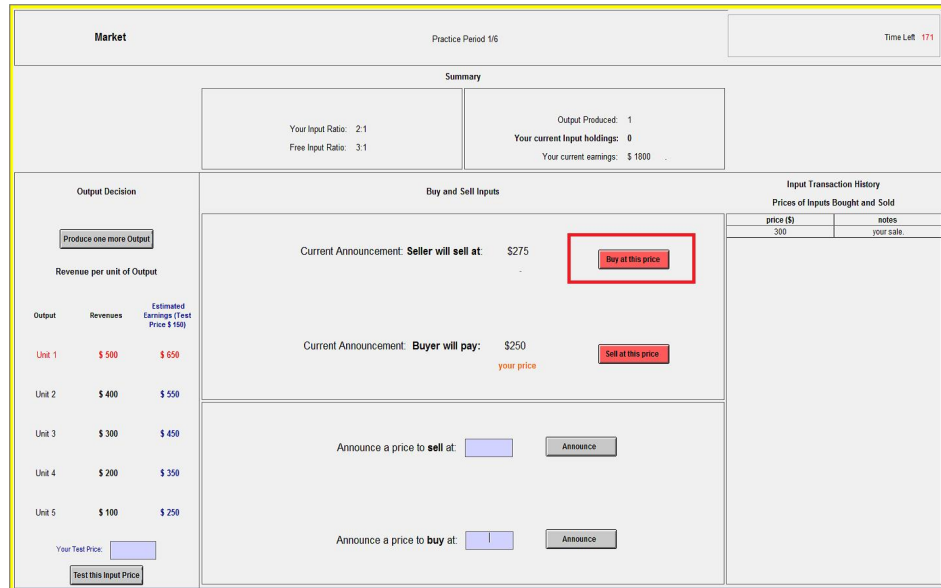


Figure 39: Market Introduction Screen 35

5.36 jan_m_36 [23s]

Notice that when you do this the transaction instantly takes place between you and the buyer who made the current announcement. The summary section at the top of the screen tells you that you now have 1 Input in your current holdings instead of 0 and that your current earnings have gone down from L\$1800

to L\$1525. Also notice that the current buying and selling announcements have been reset and the market is ready for buying and selling the next Input.

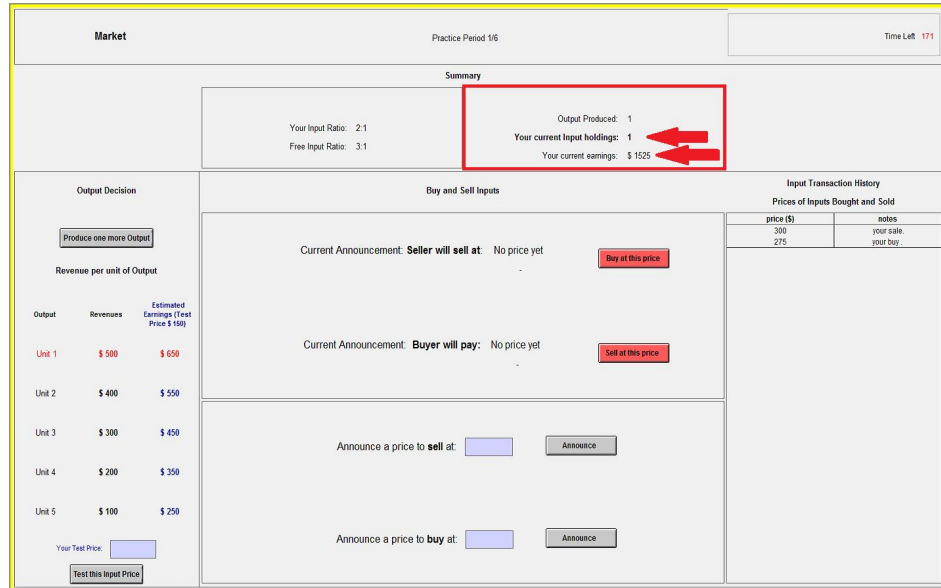


Figure 40: Market Introduction Screen 36

5.37 jan_m_37 [9]

In addition, the transaction price has again been recorded in the bottom right of your screen. Ok, that's it for the Market.

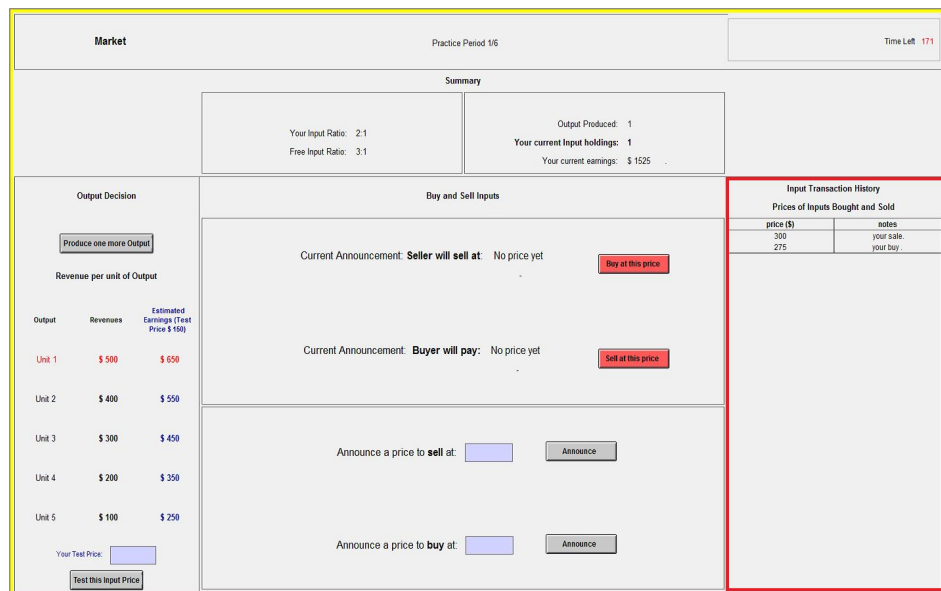


Figure 41: Market Introduction Screen 37

6 New Summary Slide

At the end of each period you will be shown a screen summarizing your actions in the period and your earnings. Take a moment to look at what information is available to you. After a few seconds this summary screen will disappear and the next period's market decision screen will be displayed. Now we will begin the six practice periods. Feel free to explore here since these periods will not affect the cash you will earn today. You will be provided with further instructions at the end of the third practice period.

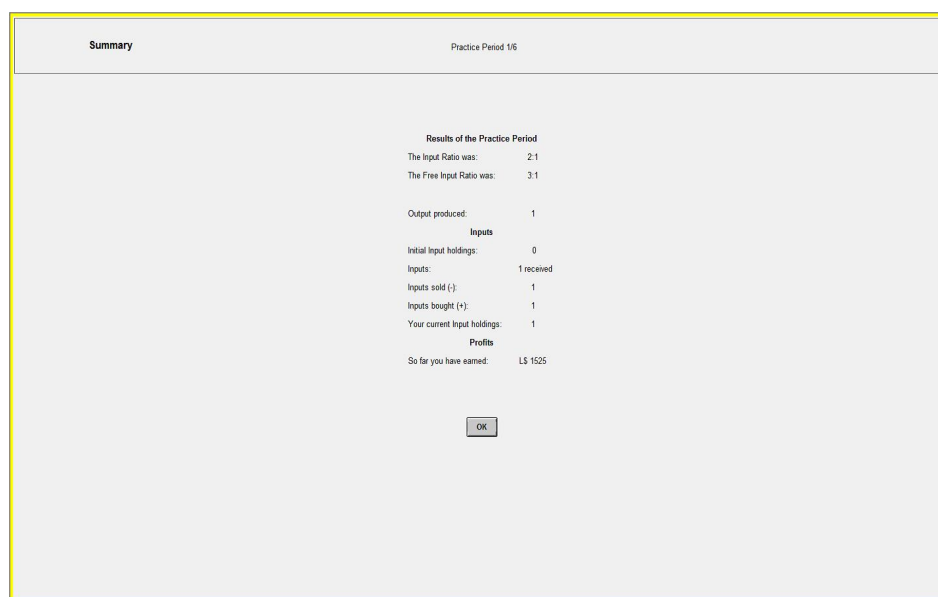


Figure 42: Market Introduction Screen 37

7 IR and Revenues 1 [TIME6]

The Input Ratio and Revenues

Now that you have seen how to produce Output and how buying and selling Inputs works we can talk more about the Input Ratio and Revenues. Starting at this point in the experiment, you will be able to choose your Input Ratio anywhere between 0 : 1 and 5 : 1 at the start of each period. If you choose an Input Ratio lower than the Free Input Ratio you will receive Inputs each time you produce Output but this technology is expensive so your revenues will be lower. If you choose an Input Ratio higher than the Free Input Ratio you will need to buy and use Inputs to produce output but this technology is cheap so your revenues will be higher.

Test question to continue (conditional OK button): If you choose the Input Ratio 4 : 1 and the Free Input Ratio is 2 : 1, in total how many Inputs will you use or receive when producing a unit of Output?

- a) use 1 unit

- b) need 1 units
- c) use 2 units
- d) need 2 units
- e) use 3 units
- f) receive 3 units
- g) use 4 units
- h) receive 4 units

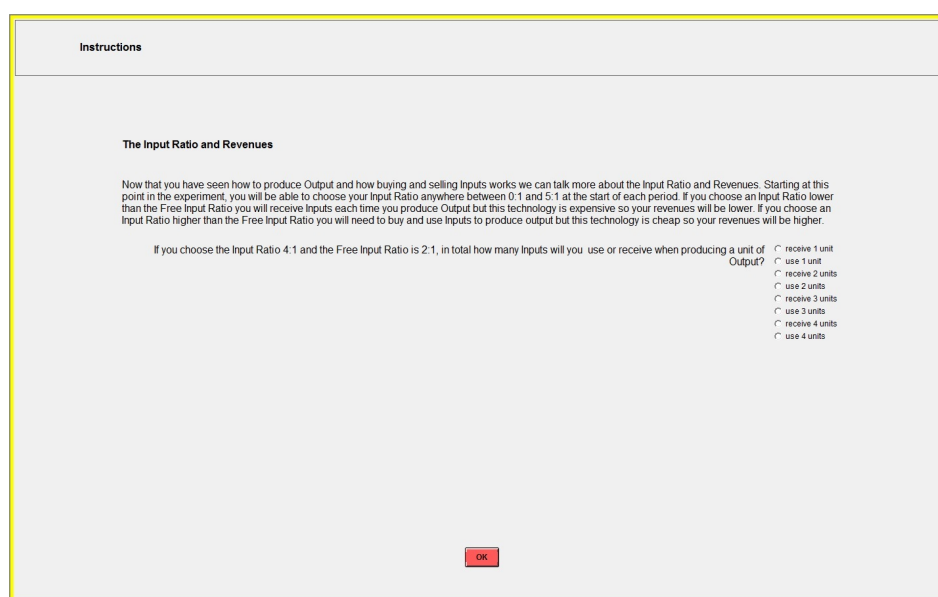


Figure 43: Market Introduction Screen 37

8 IR and Revenues 2 [TIME7]

High and Low Revenues

For the rest of the experiment there will be two revenue levels associated with each of the six different input ratios: high and low. Each period, after you have chosen your Input Ratio but before you have produced Output or bought and sold Inputs, the computer will randomly determine if everyone's revenue will be high that period or low that period. Both high and low revenue levels will be equally likely and an independent random choice will be made each period. When Revenues are high each Output produced will earn L\$60 more than when Revenues are low for all 8 participants in the experiment that period. Once everyone has selected their Input Ratio at the start of the period, everyone will then be able to produce Output and buy or sell Inputs just as we explained earlier.

The next slides will show you how you can choose your Input Ratio using information on low and high Revenues.

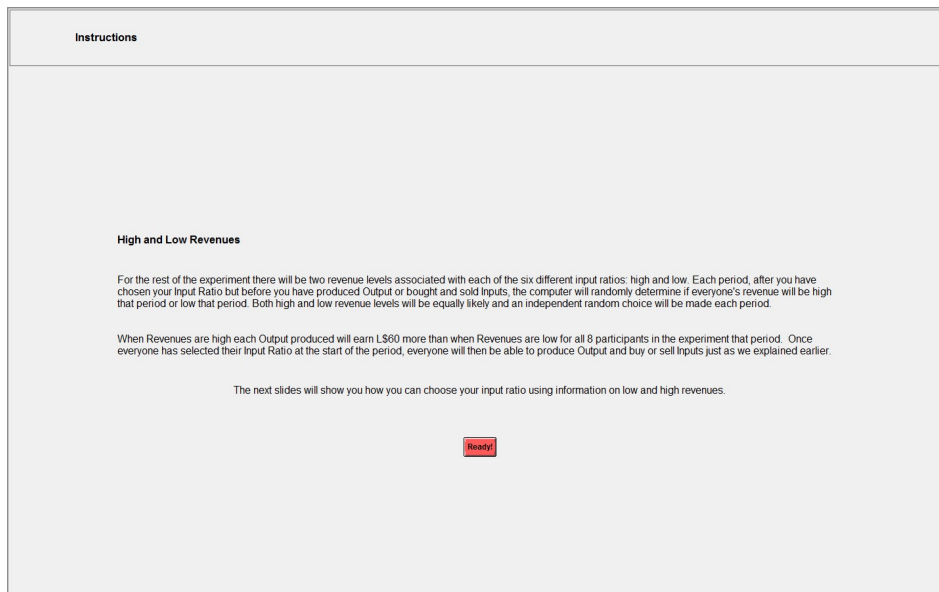


Figure 44: Market Introduction Screen 37

9 New Input Ratio Selection Slides

9.1 jan_ir_1 [19]

For each unit of Output you produce you will earn the Lab dollars shown to you in the table on the Input Ratio selection Screen. For each Input Ratio there is a column of low Revenues and a column of high Revenues. Pick any Input Ratio and any Output unit and notice that the high revenue is always \$60 higher than the low revenue.

Revenues By Intensity Ratio												Test Centre				
Output	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1		Output	Low Revenues	High Revenues	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High				
Unit 1	\$ 142	\$ 202	\$ 166	\$ 226	\$ 197	\$ 247	\$ 206	\$ 266	\$ 223	\$ 283	\$ 235	\$ 295	Unit 1	-	-	
Unit 2	\$ 112	\$ 172	\$ 136	\$ 196	\$ 157	\$ 217	\$ 176	\$ 236	\$ 193	\$ 253	\$ 205	\$ 265	Unit 2	-	-	
Unit 3	\$ 82	\$ 142	\$ 106	\$ 166	\$ 127	\$ 187	\$ 146	\$ 206	\$ 163	\$ 223	\$ 175	\$ 235	Unit 3	-	-	
Unit 4	\$ 52	\$ 112	\$ 76	\$ 136	\$ 97	\$ 157	\$ 116	\$ 176	\$ 133	\$ 193	\$ 145	\$ 205	Unit 4	-	-	
Unit 5	\$ 22	\$ 82	\$ 46	\$ 106	\$ 67	\$ 127	\$ 86	\$ 146	\$ 103	\$ 163	\$ 115	\$ 175	Unit 5	-	-	
Difference																
			-24		-21			-19				-17				-12

Your Test Price: <input type="text" value="0"/>	
<input type="button" value="Test this Ratio and Price"/>	

Free Input Ratio: 3:1

Choose your Input Ratio (Inputs:Output)

- 0:1
- 1:1
- 2:1
- 3:1
- 4:1
- 5:1

Figure 45: Intensity Ratio Introduction Screen 1

9.2 jan_ir_2 [15]

For example, for an Input Ratio of 4:1 the screen says that your 5th unit of Output will earn you L\$103 when Revenues are low and L\$163 when Revenues are high.

Input Ratio Selection													Practice Period 4/6		
Revenues By Intensity Ratio												Test Centre			
Output	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1		Output	Low Revenues	High Revenues
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High			
Unit 1	\$ 142	\$ 202	\$ 166	\$ 226	\$ 187	\$ 247	\$ 206	\$ 266	\$ 223	\$ 283	\$ 235	\$ 295	Unit 1	-	-
Unit 2	\$ 112	\$ 172	\$ 136	\$ 196	\$ 157	\$ 217	\$ 176	\$ 236	\$ 193	\$ 253	\$ 205	\$ 265	Unit 2	-	-
Unit 3	\$ 82	\$ 142	\$ 106	\$ 166	\$ 127	\$ 187	\$ 146	\$ 206	\$ 163	\$ 223	\$ 175	\$ 235	Unit 3	-	-
Unit 4	\$ 52	\$ 112	\$ 76	\$ 136	\$ 97	\$ 157	\$ 116	\$ 176	\$ 133	\$ 193	\$ 145	\$ 205	Unit 4	-	-
Unit 5	\$ 22	\$ 82	\$ 46	\$ 106	\$ 67	\$ 127	\$ 86	\$ 146	\$ 103	\$ 163	\$ 115	\$ 175	Unit 5	-	-
Difference			-24		-21		-19		-17		-12				

Your Test Price:	
<input type="text" value="0"/>	<input type="text" value="0"/>

Free Input Ratio: 3:1
 Choose your Input Ratio (Inputs:Output)

- 0:1
- 1:1
- 2:1
- 3:1
- 4:1
- 5:1

Select this Ratio

Figure 46: Intensity Ratio Introduction Screen 2

9.3 jan_ir_3 [37]

As the Input Ratio rises from 0 : 1 to 5 : 1 the revenue values associated with each ratio increases (for both low and high Revenues). This means that choosing high Input Ratios will result in higher Revenues per unit of Output. However, remember that as the Input Ratio rises you are also required to purchase more Inputs per unit of Output. The Input costs are not shown on this table and it is possible that your earnings from producing Output might go up or down as you increase your Input Ratio from 0 : 1 to 5 : 1 depending on the cost of Inputs.

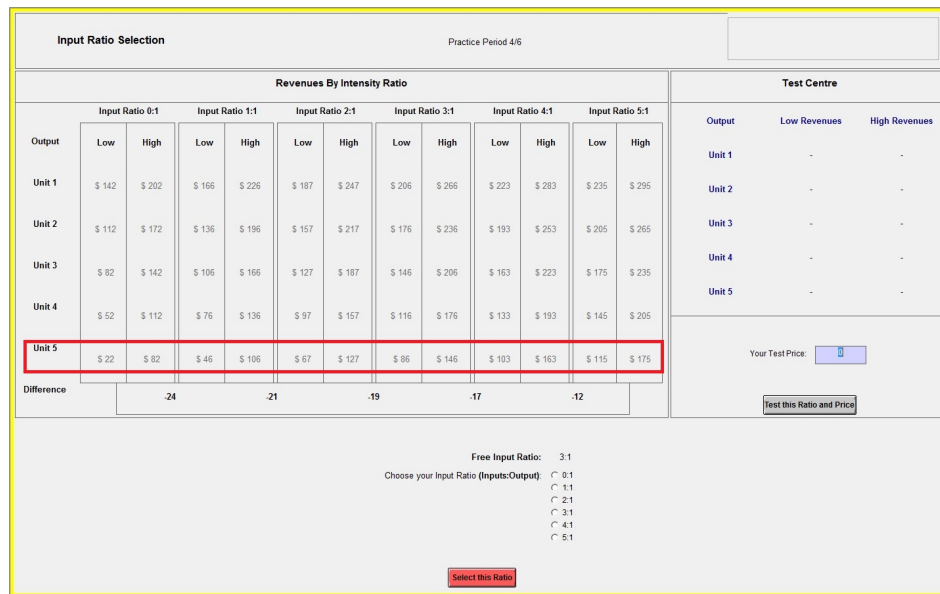


Figure 47: Intensity Ratio Introduction Screen 3

9.4 jan_ir_4 [19]

For example, notice that when your Input Ratio is 2 : 1 and Revenues are low, you will earn revenue of L\$187 for your first unit of Output. If you chose this ratio you would need 2 Inputs to produce that unit not taking the Free Input Ratio into account.

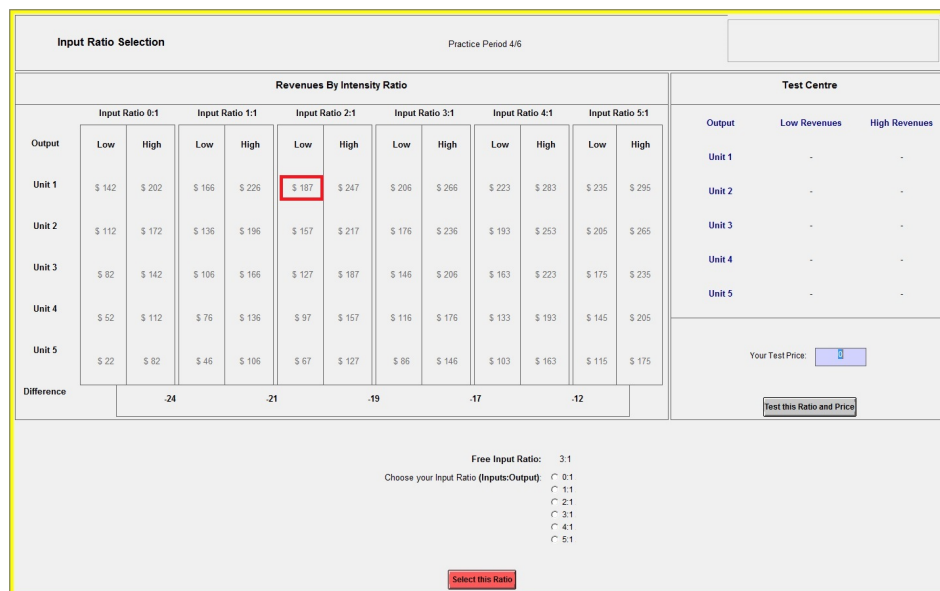


Figure 48: Intensity Ratio Introduction Screen 4

9.5 jan_ir_5 [67s]

If instead you chose the Input Ratio 3 : 1 you would earn a revenue of L\$206 for the first unit of Output if Revenues were low. Even though you would earn

L\$19 more on your first unit of Output choosing an Input Ratio of 3 : 1 instead of 2 : 1, in this case you would require three Inputs overall instead of two Inputs. Depending on what the Free Input Ratio is, choosing a higher Input Ratio will either mean you will receive less Inputs or that you will have to buy and use more inputs. Both possible consequences result in earnings that are lower than the revenue values shown because you will either have less to sell or more to buy when you choose a higher Input Ratio. If Input prices are less than L\$19 then it might be a good idea to change your Input Ratio from 2 : 1 to 3 : 1, however if the price of Inputs is greater than L\$19 it is likely not a good idea to change your Input Ratio from 2 : 1 to 3 : 1. We will discuss in a minute how you can use the 'test centre' on the right part of your screen take Input costs into account to help with your decisions.

Input Ratio Selection												Practice Period 4/6			
Revenues By Intensity Ratio											Test Centre				
Output	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1		Output	Low Revenues	High Revenues
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High			
Unit 1	\$ 142	\$ 202	\$ 166	\$ 226	\$ 187	\$ 247	\$ 206	\$ 266	\$ 223	\$ 283	\$ 235	\$ 295	Unit 1	-	-
Unit 2	\$ 112	\$ 172	\$ 136	\$ 196	\$ 157	\$ 217	\$ 176	\$ 236	\$ 193	\$ 253	\$ 205	\$ 265	Unit 2	-	-
Unit 3	\$ 82	\$ 142	\$ 106	\$ 166	\$ 127	\$ 187	\$ 146	\$ 206	\$ 163	\$ 223	\$ 175	\$ 235	Unit 3	-	-
Unit 4	\$ 52	\$ 112	\$ 76	\$ 136	\$ 97	\$ 157	\$ 116	\$ 176	\$ 133	\$ 193	\$ 145	\$ 205	Unit 4	-	-
Unit 5	\$ 22	\$ 82	\$ 46	\$ 106	\$ 67	\$ 127	\$ 86	\$ 146	\$ 103	\$ 163	\$ 115	\$ 175	Unit 5	-	-
Difference			-24		-21		-19		-17		-12				

Free Input Ratio:	3:1
Choose your Input Ratio (Inputs:Output)	<input type="radio"/> 0:1 <input type="radio"/> 1:1 <input type="radio"/> 2:1 <input checked="" type="radio"/> 3:1 <input type="radio"/> 4:1 <input type="radio"/> 5:1

Figure 49: Intensity Ratio Introduction Screen 5

9.6 jan_ir_6 [48s]

Also notice that this L\$19 difference between the Input Ratio of 2 : 1 and 3 : 1 for the first unit of Output is the same for all five units of Output going down the columns. Similar patterns exist for the other Input Ratios as well, except the revenue difference may be higher or lower. For instance, notice that with low demand and an Input Ratio of 4 : 1 the first unit revenue is L\$223, but for an input ratio of 3 : 1 the first unit revenue is 2L\$206, a L\$17 difference in revenues. In fact, for each unit of Output, revenues for an Input Ratio of 4 : 1 are L\$17 higher than for an Input Ratio of 3 : 1. For an Input Ratio of 5:1 the revenues are only L\$12 higher than those for 4 : 1.

Not recorded, added to updated file: These differences in revenues are summarized along the bottom of the Table in the row titled 'Differences'. These differences can be helpful in making your Input Ratio decision. Starting at the left of the table and moving to the right, each Input Ratio requires one more input per unit of output. Also

notice moving from left to right, that revenues per unit of output increase. You might consider moving to the right, choosing a higher Input Ratio, as long as the Revenue difference, shown in the Differences row, is greater than the price of inputs you anticipate in the market stage.

For example, suppose you expect prices to stay steady at 18. Notice that the first three differences are above 18, so you would profit moving towards the right of this table. However, moving from 3:1 to 4:1, revenues increase by only 17, therefore you should expect a loss if permit prices are 18. You might want to choose an input ratio of 3:1 if you feel strongly about your price prediction.

Input Ratio Selection Practice Period 4/6

Revenues By Intensity Ratio													Test Centre		
Output	Input Ratio 0:1		Input Ratio 1:1		Input Ratio 2:1		Input Ratio 3:1		Input Ratio 4:1		Input Ratio 5:1		Output	Low Revenues	High Revenues
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High			
Unit 1	\$ 142	\$ 202	\$ 166	\$ 226	\$ 187	\$ 247	\$ 206	\$ 266	\$ 223	\$ 283	\$ 235	\$ 295	Unit 1	-	-
Unit 2	\$ 112	\$ 172	\$ 136	\$ 196	\$ 157	\$ 217	\$ 176	\$ 236	\$ 193	\$ 253	\$ 205	\$ 265	Unit 2	-	-
Unit 3	\$ 82	\$ 142	\$ 106	\$ 166	\$ 127	\$ 187	\$ 146	\$ 206	\$ 163	\$ 223	\$ 175	\$ 235	Unit 3	-	-
Unit 4	\$ 52	\$ 112	\$ 76	\$ 136	\$ 97	\$ 157	\$ 116	\$ 176	\$ 133	\$ 193	\$ 145	\$ 205	Unit 4	-	-
Unit 5	\$ 22	\$ 82	\$ 46	\$ 106	\$ 67	\$ 127	\$ 86	\$ 146	\$ 103	\$ 163	\$ 115	\$ 175	Unit 5	-	-
Difference			-24		-21		-19		-17		-12		Your Test Price: <input type="text"/>		

Free Input Ratio: 3:1
Choose your Input Ratio (Inputs:Output):

0:1
 1:1
 2:1
 3:1
 4:1
 5:1

Figure 50: Intensity Ratio Introduction Screen 6

9.7 jan_ir_7 [23s]

At the bottom of the Input Ratio Selection Screen you will choose your Input Ratio for the period by clicking on the appropriate circle and then clicking the red button to submit your choice. The Input Ratio you choose will be private information and other participants will not be told what it is. Every period you will be able to choose your Input Ratio. You may decide to change it or leave it the same.

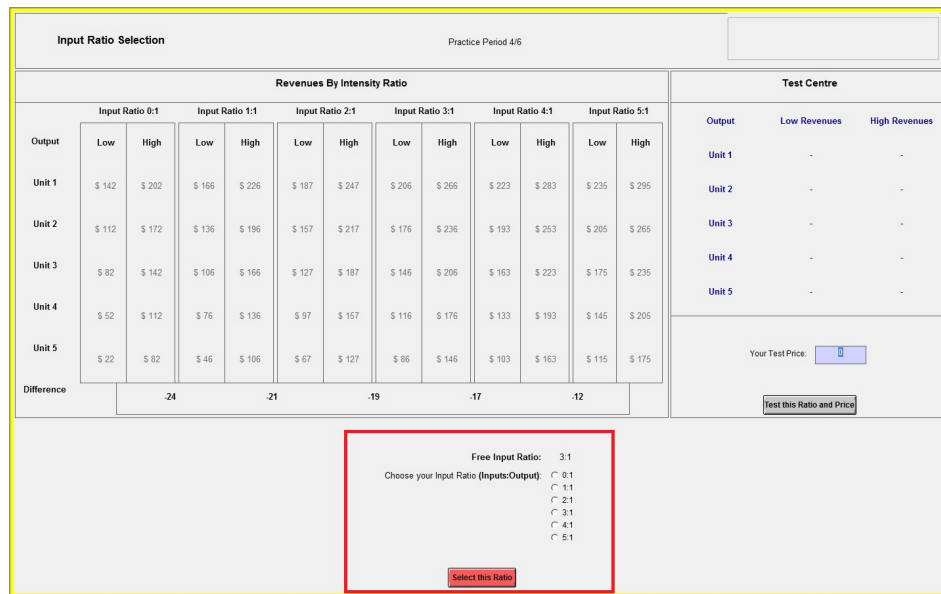


Figure 51: Intensity Ratio Introduction Screen 7

9.8 jan_ir_8 [22s]

After you have selected an Input Ratio, but before you click the red ‘Select this Ratio’ button at the bottom of the screen, you can test the Input Ratio and incorporate estimates of Input costs into your Revenues by using the ‘test centre’ at the right of the screen. For instance, if you have selected an Input Ratio of 5 : 1 and click the ‘Test this ratio and Price’ button the test centre :

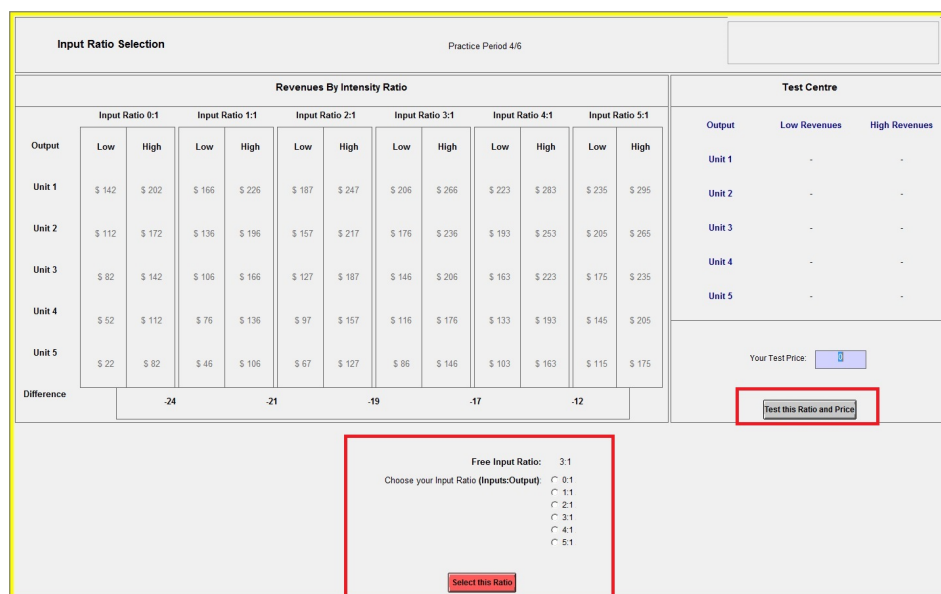


Figure 52: Intensity Ratio Introduction Screen 8

9.9 jan_ir_9 [21s]

will show the same Low and High revenue values from the main table on the left part of the screen since the Input price estimate box current is set to a price of \$L0. For instance, this means that if demand were low, the price per Input was \$L0, the Free Input Ratio 3 : 1, and you selected an Input Ratio of 5 : 1

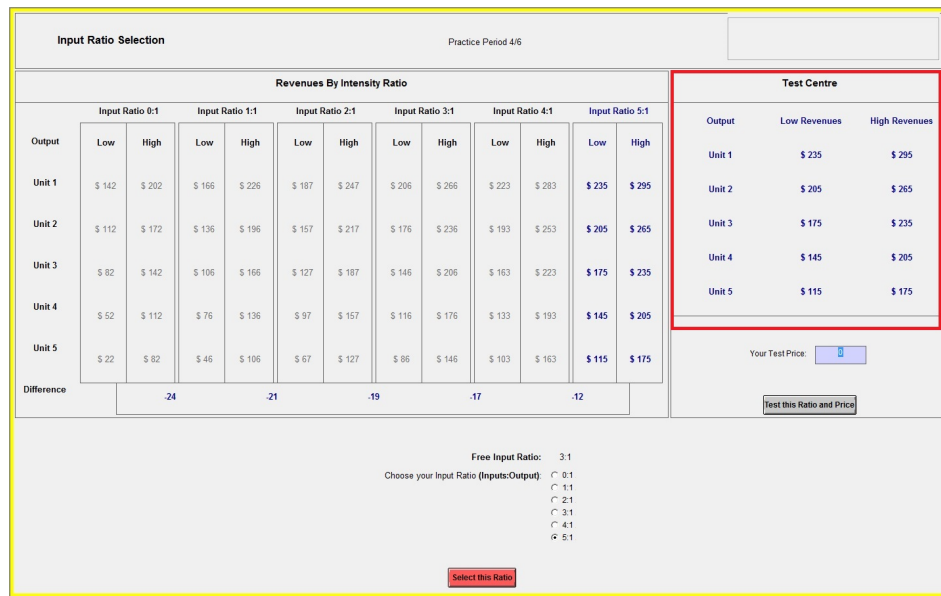


Figure 53: Intensity Ratio Introduction Screen 9

9.10 jan_ir_10 [5s]

then you would earn 235 Lab dollars for your first unit of Output,

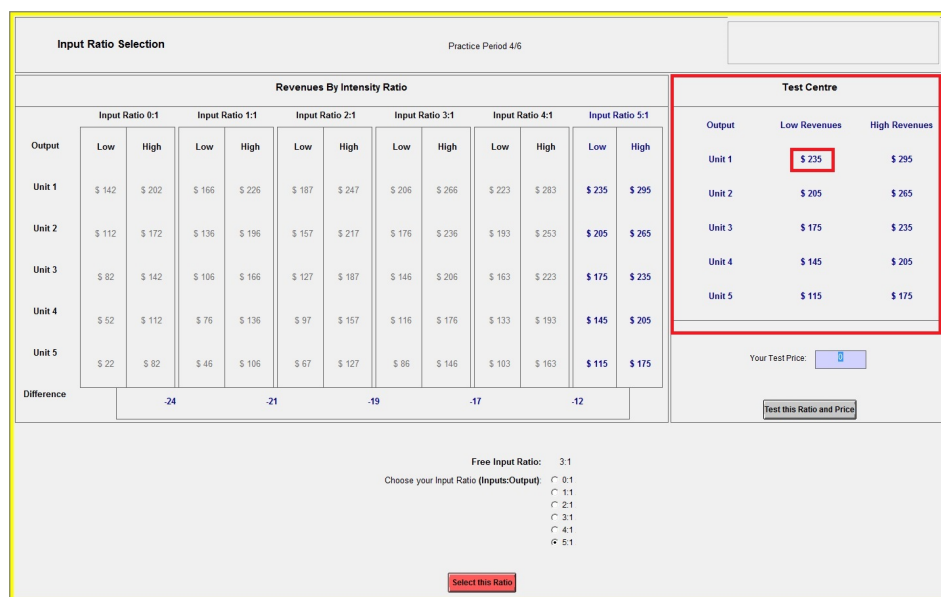


Figure 54: Intensity Ratio Introduction Screen 10

9.11 jan_ir_11 [3s]

205 for the second,

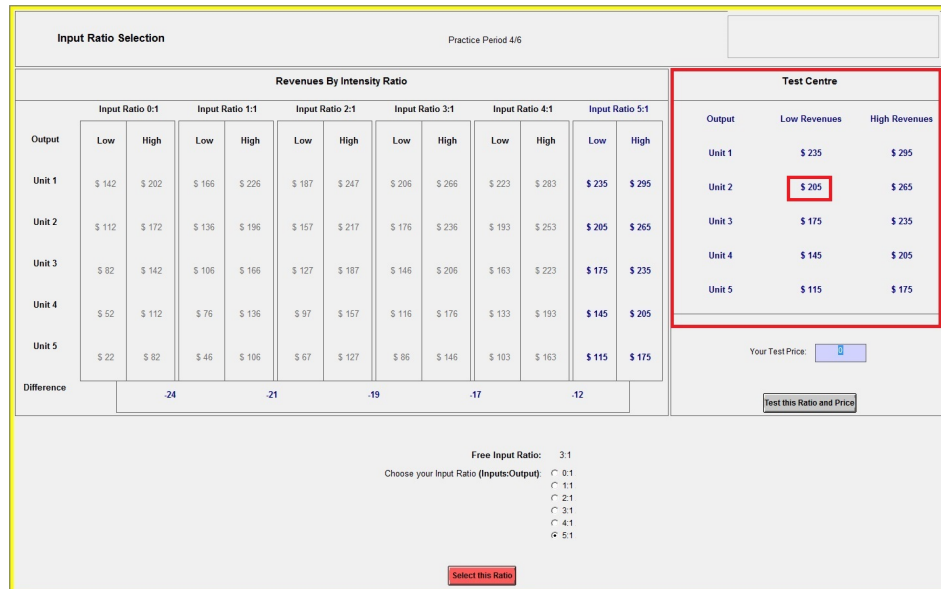


Figure 55: Intensity Ratio Introduction Screen 11

9.12 jan_ir_12 [2s]

and so on.

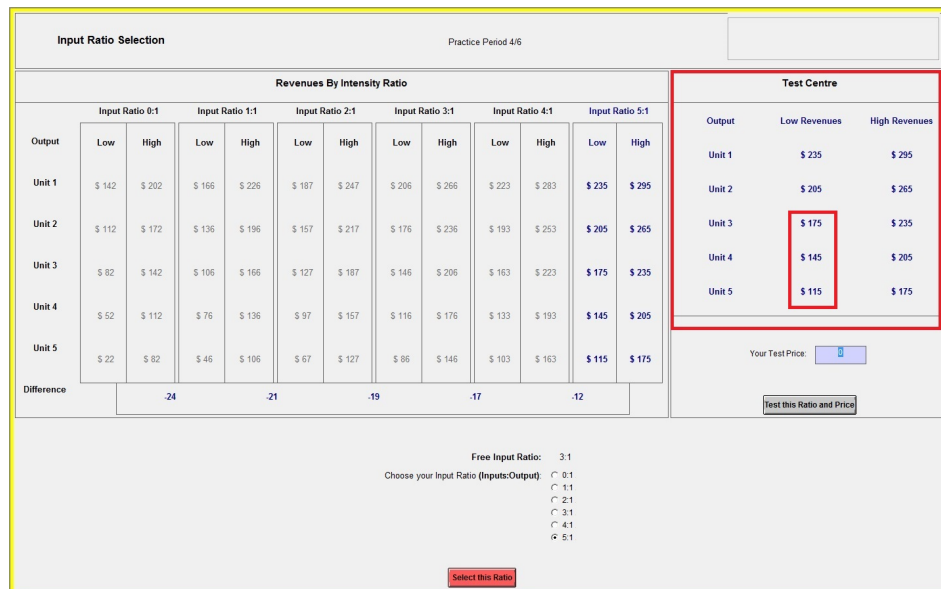


Figure 56: Intensity Ratio Introduction Screen 12

9.13 jan_ir_13 [6]

As well, you can provide the Test Centre with an estimated Input price cost to incorporate into your earnings.

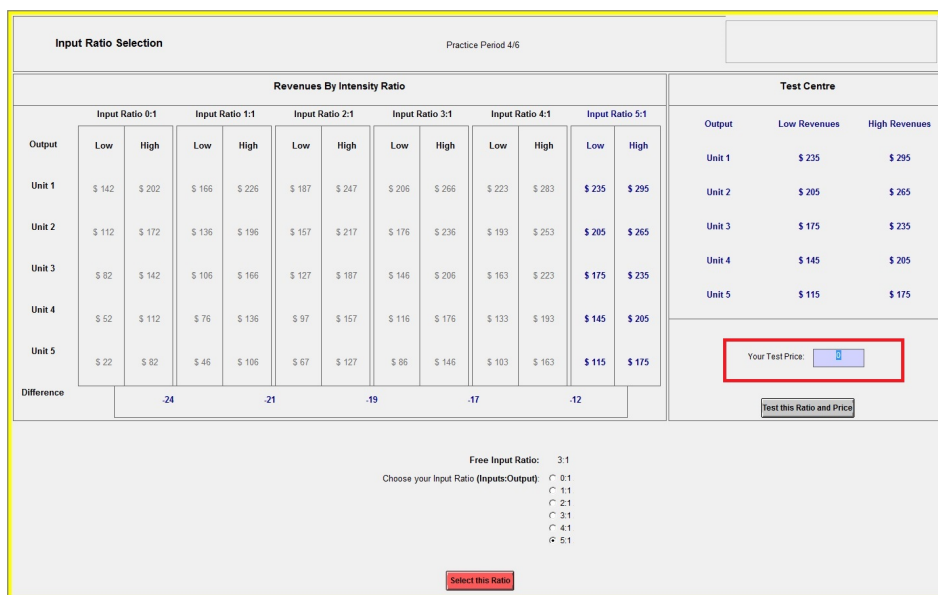


Figure 57: Intensity Ratio Introduction Screen 13

9.14 jan_ir_14 [13]

For example, suppose that Input prices have been approximately L\$25 each in the last few periods. if you were to enter an Input price estimate of L\$25 per unit of Input in the bottom right of the screen, and press the 'Test this Ratio and Price' button

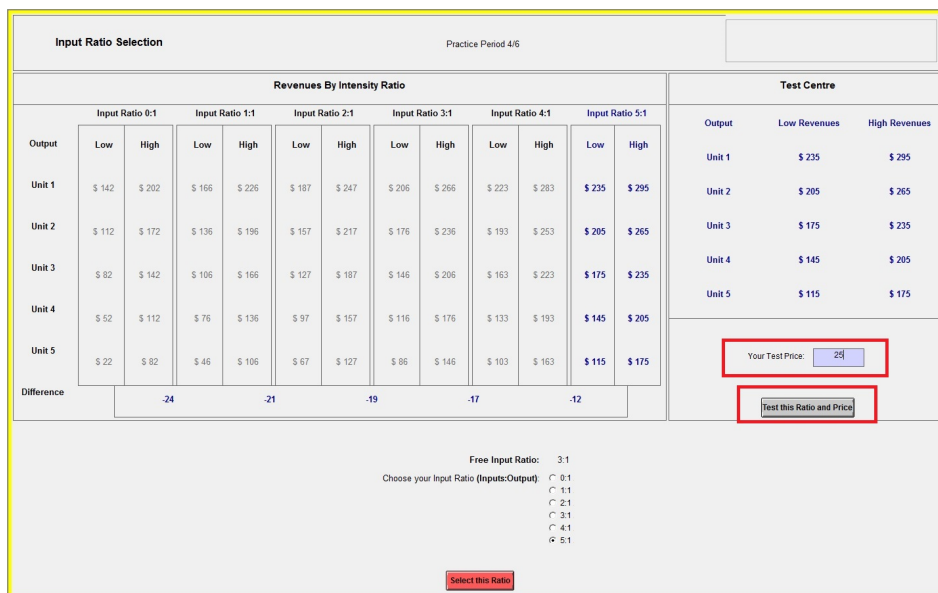


Figure 58: Intensity Ratio Introduction Screen 14

9.15 jan_ir_15 [12s]

the Test Centre will subtract the L\$25 price estimate times the 2 Inputs required to produce each unit of Output - so L\$50 is subtracted from the revenue of each

Output you could produce.

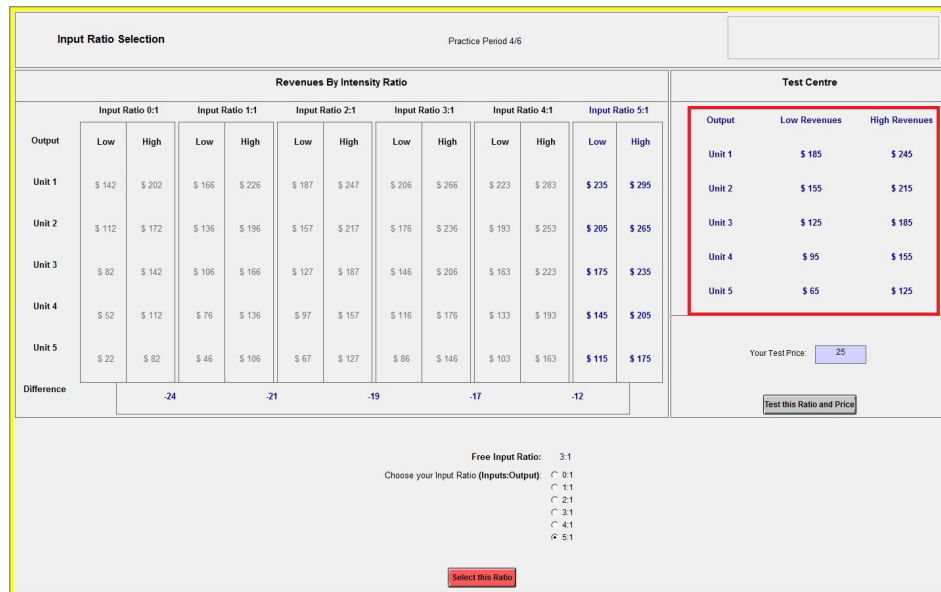


Figure 59: Intensity Ratio Introduction Screen 15

9.16 jan_ir_16 [13s]

Notice now with low demand that your earnings after receiving Revenues and paying these estimated Input costs would be 185 Lab dollars for your first unit of Output, L\$155 for the second, and so on.

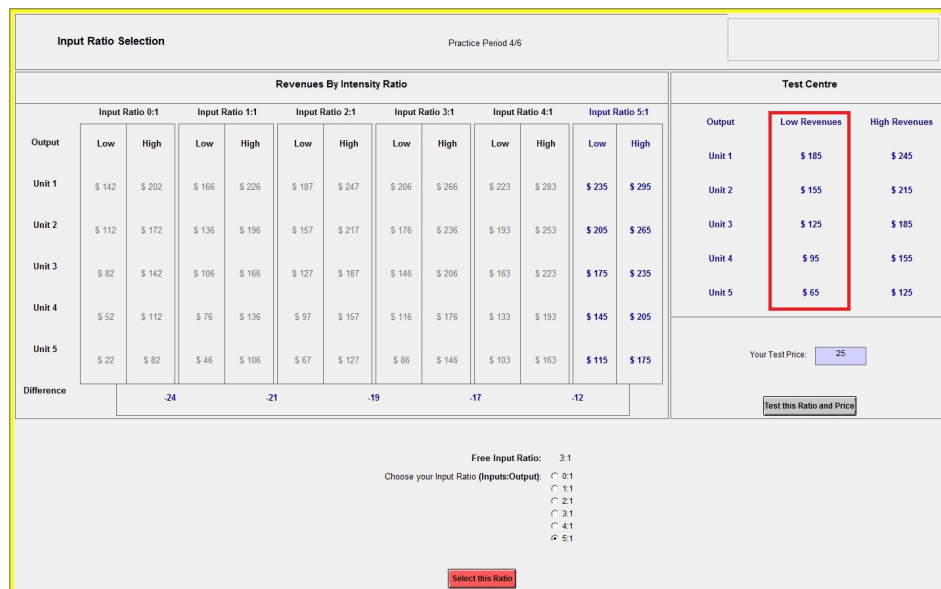


Figure 60: Intensity Ratio Introduction Screen 16

10 Begin IR Practice [TIME9]

After you select an Input Ratio for the period you will have a chance to use the market screen to produce Outputs and buy and sell Inputs just like you did in the first three practice periods. Getting a hang of this Input Ratio choice stage will take some time so we will give you another 3 unpaid practice periods to try things out with a new set of revenue values that are different from your first three practice periods. Be aware that the revenue values on your decision screen might be different from those you have just seen in these instructions. For practice periods 4 to 6 the Free Input Ratio is 3 : 1, which will be displayed to you on the lower portion of the Input Ratio selection screen. Everyone will have this same Free Input Ratio for each of the next three practice periods. The next three rounds are not paid, so feel free to experiment. Other participants may have different revenue values than you do. Before the paid part of the experiment begins we will pause and make sure you are ready to continue. Please raise your hand now if you have any questions.

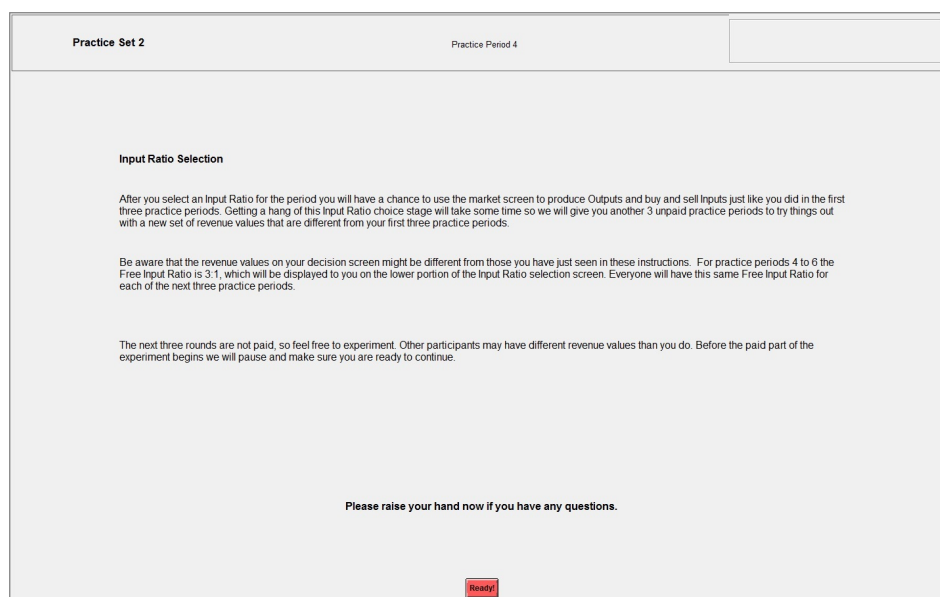


Figure 61: Intensity Ratio Screen: Begin second set of practice periods

11 Notice of Start of Paid Periods [TIME10]

Now that you have completed the practice periods, we are ready to begin the paid periods of the experiment. The experiment will reset again now. Your revenues will be different from those you saw in the practice periods but will remain constant for all paid periods. Other participants may have different revenue values than you do. For all of the paid periods everyone in the experiment will have a Free Input Ratio of 2 : 1.

For the first 7 Paid periods the Market will last 3 minutes (180 seconds), after that the Market will last 2 minutes (120 seconds).

The experiment will have a minimum of 10 paid periods. After the tenth period there is a 1 in 6 chance that the experiment will end. If the experiment does not end we will continue for another paid period. At the end of the 11th paid period (if there is one) there will again be a 1 in 6 chance that the experiment will end and hence a 5 in 6 chance that the experiment will continue for another paid period. This process has been programmed into the experiment, and will repeat at the end of each additional period until the experiment ends. Therefore the experiment could end at the end of the 10th period, but it is more likely that the experiment will last longer than that.

End of the Experiment

At the end of the experiment, any unused Inputs will be worth nothing. After the last Period ends the computer will show you the lab dollar earnings that you accumulated from the first paid period to the last paid period and the conversion to Canadian dollars. You will also be shown the results of the individual task you completed at the beginning of today's session. When you click the 'OK' button you will be asked to complete a few optional demographic questions. Your earnings today will not be revealed to other participants so shortly after completing the demographic questions you will be called one by one into the lab manager office to privately collect your cash. No communication between participants is allowed at any time during the experiment. Please raise your hand at any time if you have any questions.

You have completed all the practice rounds. The first paid period of the experiment will begin after everyone has pressed the 'Ready!' button below.

Please raise your hand now if you have any questions. After any questions have been addressed the experiment will begin.

Start of Paid Periods
Paid Period 1

Paid Periods

Now that you have completed the practice periods, we are ready to begin the paid periods of the experiment. The experiment will reset again now. Your revenues will be different from those you saw in the practice periods but will remain constant for all paid periods. Other participants may have different revenue values than you do. For all of the paid periods everyone in the experiment will have a Free Input Ratio of 2:1.

For the first 7 Paid periods the Market will last 3 minutes (180 seconds), after that the Market will last 2 minutes (120 seconds).

The experiment will have a minimum of 10 paid periods. After the tenth period there is a 1 in 6 chance that the experiment will end. If the experiment does not end we will continue for another paid period. At the end of the 11th paid period (if there is one) there will again be a 1 in 6 chance that the experiment will end and hence a 5 in 6 chance that the experiment will continue for another paid period. This process has been programmed into the experiment, and will repeat at the end of each additional period until the experiment ends. Therefore the experiment could end at the end of the 10th period, but it is more likely that the experiment will last longer than that.

End of Experiment

At the end of the experiment, any unused Inputs will be worth nothing. After the last Period ends the computer will show you the lab dollar earnings that you accumulated from the first paid period to the last paid period and the conversion to Canadian dollars. You will also be shown the results of the individual task you completed at the beginning of today's session. When you click the "OK" button you will be asked to complete a few optional demographic questions. Your earnings today will not be revealed to other participants so shortly after completing the demographic questions you will be called one by one into the lab manager office to privately collect your cash. No communication between participants is allowed at any time during the experiment. Please raise your hand at any time if you have any questions.

You have completed all the practice rounds. The first paid period of the experiment will begin after everyone has pressed the Ready! button below.

Please raise your hand now if you have any questions. After any questions have been addressed the experiment it will begin.

Ready!

Figure 62: Intensity Ratio Screen: Begin second set of practice periods

Discussion and conclusions

This thesis explored applications and methods for confronting theory with evidence. Each chapter took a scientific approach to investigating the match between a well-defined theory and carefully collected evidence. Chapters 1 and 2 developed new frameworks to evaluate the match between evidence and established theoretical propositions and guidelines. In Chapter 3 new evidence was generated to offer deeper insights into the relative performance of Cap and Trade and Intensity Target market programs for control of pollutants. The main contribution of this work is to improve the ability to answer the important question of whether or not a particular set of data is reflective of pre-defined relationship. In doing so, this thesis highlights the scientific nature of economics.

In Chapter 1 the evidence consisted of observations from an economic experiment and the analysis relied upon nonparametric estimation. The combination of experimental data and this particular analytical method was exceptionally well suited to the experimental economist's scientific pursuit of truth. In experimental economics a heavy emphasis is placed on the design of the experimental environment. This means that the data collected in economics experiments reflect more controlled interactions and the processes of their generation can be better understood relative to the empirical alternatives such as population surveys and administrative data sets. Analysis of experimental data often favours simple analytical methods based on evalua-

tion of treatment effects, but, at times, Standard econometric regression techniques are used. Standard techniques are based in estimating parameters of an unknown population, relying upon strong assumptions to do so. Chapter 1 highlights the substantial gains to be made from adopting nonparametric regression techniques in instances when regression analysis is desired. Because experimental data sets are composed of carefully collected observations explicitly generated to understand specific behavioural responses to changes in the economic environment, not as an attempt to estimate parameters for a larger population, the Nonparametric approach offers a satisfying sense of coherence between data collection and analysis and provide a stronger ability to reveal relationships within the data. The method presented in Chapter 1 is in no way limited to experimental data applications. The method provides an improvement in the ability to analyse the match of data sets featuring a discrete outcome ranging over a continuous covariate, with theoretical propositions, regulations or guidelines which posit single discrete changes. Chapter 2 adapted and expanded the basic analytical framework of Chapter 1 to confront a clinical practice guideline with administrative data. In this case the framework serves to improve the ability to answer the question of whether or not medical practitioners adhere to a practice guideline. The flexibility of the framework means that there is substantial value to be gained from integration of this framework into a broader system of health practitioner assessment. At the same time, the framework is generalizable to other settings involving evaluation of regulations, guidelines and theory. The second chapter represents a substantial expansion upon the first chapter on a number of fronts. The first is to extend the methodological framework to the setting of evaluation of health care provider adherence to clinical practice guidelines. The second is to incorporate identification of multiple discrete changes within a set of data. The third is to improve the conversion of exist-

ing health data into useful information concerning health care practitioner performance. Lastly, the framework developed in this chapter is complementary to the system of appropriateness of care detailed by Brook (2009) as it contributes a uniform method to gauge guideline adherence which is comparable across health care sectors and regions.

Chapter 3 confronted theory about the performance of emission permit trading markets with evidence generated in the experimental economics laboratory. Industrial emissions and emission control have become important and controversial topics of global interest. The comparison of alternative market based trading mechanisms offers important insights into the field implementability of each. This work was built upon existing experimental frameworks, extending emission permit trading experiments to a more realistic setting. In light of recent agreements between the US and Canada to reduce emissions of greenhouse gases, this work serves to offer a potentially viable alternative to direct taxes or Cap and Trade programs. In this case theory suggested that an Intensity Target emission trading program would be an inferior choice for attaining a given level of pollution. Confronting this theory with experimental data, however, led to the interesting finding that Intensity Target programs may pull an economy towards cleaner production technologies while maintaining high levels of efficiency. The overall efficiency comparison between Intensity Targets and Cap and Trade depends on the exact curvature of the environmental damage function which, in turn, depends on the nature of the pollutant that is being regulated. The value of experimental economic approaches in providing evidence to better answer questions about theoretical propositions is illustrated in this chapter.

This thesis shows that by combining available analytical techniques and gathering good evidence, economics is an invaluable scientific tool in assessing a wide range of interactions. Ultimately, improvements in assessment and in-

sight provided by this approach can support the development of innovative solutions to a multitude of problems facing humanity.

References

- Brook, R. H. (2009). "Assessing the appropriateness of care - its time has come".
In: *Journal of the American Medical Association* 302.9, pp. 997–998.

