DP2009/19

Whatever next? Export market choices of New Zealand firms

Richard Fabling, Arthur Grimes
and Lynda Sanderson

December 2009

JEL classification: D21, F10, L25

www.rbnz.govt.nz/research/discusspapers/

Discussion Paper Series

ISSN 1177-7567

DP2009/19

Whatever next? Export market choices of New Zealand firms[*]

Richard Fabling, Arthur Grimes and Lynda Sanderson[†]

Abstract

We examine product and market entry choices of New Zealand exporters, using an enterprise level dataset which links firm performance measures with detailed data on merchandise trade. We focus our enquiry not on the broad question of what determines a firm's ability to export, but on the subsequent question: given that a firm has the ability to export, what determines the choices they make about what and where to export?

We simultaneously consider firm and market level determinants of export market entry. At the firm level we find that measures of general and specific prior trade experience play an important role in determining the firm's future export activities. That is, we find evidence of path dependence within firms. We also find evidence of path dependence across firms, with entry into new export relationships reflecting demonstration effects from the export activities of other firms in the local area. These results are robust to the inclusion of other determinants of exporting, including the macroeconomic performance of destination countries, exchange rate movements, and the past performance of the exporting firm.

[*] The Reserve Bank of New Zealand's discussion paper series is externally refereed. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Reserve Bank of New Zealand. The authors wish to thank Statistics New Zealand for access to the data, Jacques Poot for discussions on rare event models, and Motu Economic and Public Policy Research seminar participants and an anonymous referee for helpful feedback.

[†] *Richard Fabling:* RBNZ (richard.fabling@rbnz.govt.nz); Motu Economic and Public Policy Research. *Arthur Grimes:* Motu Economic and Public Policy Research (arthur.grimes@motu.org.nz); University of Waikato. *Lynda Sanderson:* RBNZ (lynda.sanderson@rbnz.govt.nz); University of Waikato.

ISSN 1177-7567

Disclaimer

This research uses data that was accessed while Richard Fabling and Lynda Sanderson were on secondment to Statistics New Zealand in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Act are allowed to see data about a particular business or organisation. The results of this work have been confidentialised to protect individual businesses from identification. The analysis and interpretation of these results were undertaken while Richard Fabling and Lynda Sanderson were at the Reserve Bank of New Zealand and Arthur Grimes was at Motu. The opinions, findings, recommendations and conclusions expressed in this report are those of the authors. Statistics New Zealand, the Reserve Bank of New Zealand, Motu and the University of Waikato take no responsibility for any omissions or errors in the information contained here.

The results are based in part on tax data supplied by Inland Revenue to Statistics New Zealand under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information is published or disclosed in any other form, or provided back to Inland Revenue for administrative or regulatory purposes. Any person who had access to the unit-record data has certified that they have been shown, have read and have understood section 81 of the Tax Administration Act 1994, which relates to privacy and confidentiality. Any discussion of data limitations or weaknesses is not related to the data's ability to support Inland Revenue's core operational requirements.

Statistics New Zealand protocols were applied to the data sourced from the New Zealand Customs Service. Any discussion of data limitations is not related to the data's ability to support that agency's core operational requirements.

1 Introduction

The international literature provides broad support for the assumption that sunk costs influence firms' export decisions. However, until recently firm-level research in this area has tended to treat export status as a binary variable – firms are either exporting or they are not. Hence empirical studies of entry into exporting have focused on the initial entry decision, particularly on identifying the firm-specific characteristics which set exporting firms apart from non-exporters. We focus our enquiry on a subsequent question: Given that a firm *has* the ability to export, what determines the choices they make about what and where to export?

Focusing on the behaviour of already-exporting firms is essential for understanding the processes by which aggregate export value increases over time. Fabling and Sanderson (2010) document that a large proportion of aggregate trade growth in New Zealand over the past decade has come from expansion in the range of export activities undertaken by incumbent exporters. These firms account for over four fifths of net growth in the annual average value of merchandise trade between 1996-98 and 2004-06. In turn Fabling and Sanderson (2010) show that over two thirds of that growth was created by incumbent exporters entering into new trade relationships. This effect dwarfs the impact of firms' initial export entry in terms of material effect on overall export growth.[1]

The literature points to the importance of sunk costs in determining firms' initial export entry decisions. At least theoretically, this argument seems equally persuasive for subsequent entries. Every geographic or product market provides new challenges for firms, including setting up distribution networks, and coming to grips with foreign consumer preferences and government regulations. However, firms may become more adept at handling these challenges over time, building up both market-specific knowledge and networks, and general exporting competencies.

To identify the existence of relationship-specific sunk export costs, we look at whether firms' past experience of exporting influences the choices they make about entry into new trade relationships – once a firm has exported a product to one country, is it more likely to send the same product to other destinations? Does an existing trade relationship tend to increase the

[1] In a similar decomposition, Bernard et al (2009) find that changes in the product-country mix of existing exporters account for 42 percent of net export growth among US exporters between 1993 and 2003, well above the share associated with net export entry and exit (24 percent) or net growth in existing relationships (35 percent).

probability that new products will be exported to the same country?[2] Finally, we consider whether one firm entering a new export market creates spillover benefits to other firms by providing an example which they can follow.

These questions are examined using firm-level longitudinal trade and performance data for New Zealand. Existing New Zealand and international literature shows that high performing firms self-select into exporting (Wagner (2007) provides a recent review). We include firm performance variables to test whether this is also true for subsequent entry events.[3] We also include variables reflecting the incentives to enter specific markets, such as the size, wealth and openness of potential trade partners and the relative exchange rate. Finally, to reflect differences in the sunk costs of entry into new trade relationships, we include the firm's own history of international engagement and variables measuring demonstration effects from other exporting firms.

Section 2 describes our conceptual model, drawing on the existing literature on export market entry. Section 3 outlines the data source, sampling strategy, and explanatory variables, while section 4 outlines the estimation approach. Sections 5 and 6 discuss the main empirical results and robustness tests respectively. Section 7 concludes.

2 Conceptual framework

We consider the determinants of entry into new trade relationships, where a relationship is defined as a firm exporting a specific product to a specific destination. As such, a new entry may involve the export of an existing product to a new market, a new product to an existing market, a new product to a new market or a new combination of existing products and markets.

Several papers have considered the relationship between export diversification, characteristics of the firm, and the wider economic environment. One strand of this literature focuses on the product dimension. Bernard et al (2006) develop a model in which firm-level product diversity is driven by a combination of firm productivity (affecting production costs of all products) and a stochastic firm-product-country level "consumer tastes" draw

[2] Fabling and Sanderson (2010) show that this type of incremental expansion is a key source of export growth, with the export of new or existing product lines to existing trade partners accounting for nearly half of aggregate export value growth.

[3] More precisely, we test whether exporters that enter additional markets have higher initial productivity than exporters that don't enter additional markets.

(determining the destination-specific popularity of each of the firm's potential product lines). They show that a fall in trade costs (including tariffs) will lead firms to concentrate their efforts on a smaller number of core products. However, while the model generates empirically supportable predictions regarding product diversity, the nature of the consumer tastes variable means that actual outcomes for any given product-country relationship are random.[4] Eckel and Neary (2009) also consider the impact of trade liberalisation on product diversification, but in a model in which firms face increasing production costs as they move further away from their "core competencies". Again, their model predicts greater diversification among high productivity firms and a narrowing of focus in response to trade liberalisation.

A second strand of the literature focuses on the geographic dimension. Authors such as Eaton et al (2008) and Melitz and Ottaviano (2008) posit that incentives to export are determined by destination market characteristics (eg, market size and distance from the home market) as well as firm characteristics. These models imply a hierarchy of potential destinations in which low productivity firms choose to enter only the easier or more attractive markets while more productive firms export to a wider range of destinations.[5] In both cases, however, market entry costs are assumed to be exogenous to the individual firm.

Our focus is on "learning to export" – that is, on the relationship between past international experience and entry costs for new relationships. To our knowledge this paper is the first to simultaneously consider the product and destination dimensions of firm-level export relationships in a model of endogenous sunk costs.

In this way our work contrasts with existing literature which has focused more directly on the estimation of sunk costs, such as Das et al (2007) who consider firm-level data but only with respect to the initial entry into exporting, Bernard and Jensen (2004) who consider the degree of export hysteresis at the firm level, and Rauch (1999), who considers the product dimension, focusing on the importance of proximity and linguistic similarity in determining relative export propensities for differentiated and undifferentiated products.

Our method builds on the fact that exporting incurs many costs. Some of these are variable costs, including transport, insurance and tariffs, which

[4] The authors note the possibility of imposing greater structure on the consumer tastes draw to reflect correlations in tastes across countries and products.

[5] An alternative model (Baldwin and Harrigan 2007) suggests that the source of firm heterogeneity is differences in product quality and that only those firms which produce high quality goods will enter more "difficult" markets.

lower the value of each unit of exports to the firm. Others are fixed but incurred on an ongoing basis, such as costs for maintaining offshore sales offices or ongoing relationships with distributors. Finally, firms face sunk costs associated with entry into new markets, including information costs such as market research on the structure of demand in foreign markets, setting up distribution networks, and learning about the regulations and institutional requirements of foreign markets. Fixed costs of export market entry are generally believed to be significant relative to marginal shipping costs.[6]

Although geographic market entry costs are a more common feature of export theories, firms also incur costs from entry into new product markets. These include the direct costs of developing a new product but also many costs associated with market entry, such as identifying market demands and tailoring marketing strategies to encompass new products.[7]

The empirical specification of our paper is similar to that of Evenett and Venables (2002), who use aggregate export data by three-digit product and destination for a panel of 23 developing and middle income countries to examine what they call "the geographic spread of trade" – the export of existing product lines to new trading partners. They find evidence that geographic and linguistic proximity to both the home market and existing export destinations play a role in determining the probability of expansion into previously unsupplied markets, implying a role for learning from existing export experiences. However, the use of product line data prevents identification of the micro-economic channels underlying this pattern.

The key assumption of sunk market entry costs suggests a number of testable hypotheses, many of which have been addressed in the literature to date. Here, we recap hypotheses associated with initial export entry and extend them to cover entry into additional markets and products.

[6] Das et al (2007) find that while initial entry costs are high, per period continuation costs are negligible on average, but important for at least some firms. Anderson and van Wincoop (2004) consider various sources of trade costs, including both marginal, volume-related costs and information, language and distribution costs. They find that policy barriers (tariffs and non-tariff barriers) add around eight percent to the cost of getting exports to consumers, with transportation costs (including monetary and time costs) adding 20 percent. Overall, they suggest that total trade costs create an effective *ad valorem* tax averaging around 170 percent for industrialised countries. These estimated trade costs differ dramatically across countries and products.

[7] As the data we have available does not include product level information on firms' domestic sales we cannot distinguish between existing product lines which are being newly exported and new product lines which are exported as soon as they are developed. If the latter situation dominates, the implied cost of export entry may be overstated as it will reflect both development and export-related costs.

Consider the model developed by Clerides et al (1998) where "incumbent exporters continue to export whenever current net operating profits plus the expected discounted future payoff from remaining in exporting is positive, and non-exporters begin to export whenever this sum, net of start-up costs, is positive. Expected future payoffs include the value of avoiding start-up costs next period and any positive learning effects that accrue from foreign market experience."

More formally, define y_t as a dummy variable indicating whether a firm exports in the current period ($y_t = 1$), or not ($y_t = 0$); $\pi^f(c_t, z_t^f)$ as the profit available from foreign markets, given marginal cost c_t (assumed to be constant across units within any given time period) and the current conditions in foreign markets z_t^f; M_t as the per period fixed cost of being an exporter (eg, costs of dealing with intermediaries); $\delta[E_t(V_{t+1}|y_t = 1)] - E_t(V_{t+1}|y_t = 0)]$ as the expected future value in the next period, conditional on being an exporter in the current period, less the expected future value in the next period conditional on not being an exporter in the current period, all discounted by the one-period discount factor δ; and F as the fixed cost of market entry, incurred only when the firm was not exporting in the previous period ($y_{t-1} = 0$). Firms export whenever

$$\pi^f(c_t, z_t^f) - M_t + \delta[E_t(V_{t+1}|y_t = 1)] - E_t(V_{t+1}|y_t = 0)] \geq F(1 - y_{t-1}). \quad (1)$$

For export relationships, rather than a binary export decision, we must add country and product subscripts to each of the relevant variables. Consider a firm deciding whether to export for the first time or an incumbent exporter deciding whether to export a new product or enter a new geographic market. These decisions are effectively identical to that proposed by Clerides et al, with the addition that firms must choose which market(s) is likely to provide the best returns and, for multi-product firms, whether to export all or only part of their range.[8] Each geographic or product market entry involves additional fixed costs. However, firms may be able to gain economies of scope by entering into multiple relationships. For example, by exporting multiple products to a single country firms incur additional development and marketing costs for each new product but can spread the costs of learning about institutional settings across a wider range of goods.

[8] While we leave open the possibility that export experience affects marginal production costs, we focus on the potential impact on fixed costs of additional market entry. Fabling and Sanderson (2009) examine the impact of exporting on firm performance in New Zealand, finding that export entry has a causal effect on total employment but not on productivity. In support of the empirical results presented in this paper, they also find a strong self-selection effect for entering exporters.

Other factors which may lower relationship-specific entry costs include experience with other forms of international engagement, such as FDI, joint ventures, offshore production or direct imports, or demonstration effects from the export activities of other firms. Country- and relationship-specific import experience may reduce costs of market entry as a firm may already have some knowledge about conditions in the destination country. Past importing of a product may be important if firms are able to learn to produce a new variety by copying from an established offshore producer or if some portion of their export activities are actually in re-exports.[9]

As well as learning by experience, firms may also be able to learn from the experiences of others. Hausmann and Rodrik (2003) discuss the role of demonstration effects in allowing firms to recognise market opportunities. In their model, entrepreneurial behaviour is limited by the inherent risks associated with innovating. Firms may observe their competitors moving into new markets and follow suit, allowing them to better choose markets, reduce the risks associated with entry, and (potentially) bid away the rents accruing to the first mover. At the same time, demonstration effects may help firms to directly reduce the costs of market entry, through easier access to information and networks needed to smooth their entry into that market.

Research looking at firms' overall export propensity (the probability of entering their first export relationship) has tended to find little evidence for export demonstration effects,[10] though there have been some exceptions. For example, Greenaway and Kneller (2004) find consistently positive export propensity spillovers and that a large number of new entries to the export market have a greater effect than a high concentration of existing exporters.[11]

Finally, economic conditions both at home and abroad may impact on both the decision to export and which countries to target. Early studies of New Zealand export behaviour found that changes in manufacturing exports could be explained in a large part by domestic GDP – when domestic incomes were low, exports rose as firms sought new outlets for their output (Morgan

[9] That is, if some of their export products are brought in from offshore, undergo minor alterations (eg, repairs, repackaging), and are then re-exported.

[10] For example, Aitken et al (1997) and Bernard and Jensen (2004).

[11] This result may be driven in part by changing macroeconomic conditions either domestically or abroad. For example, if some firms are slower to react to new export incentives than others, the laggards will look like they have been influenced by the early entrants. To mitigate this issue we include explicit controls for macro-economic conditions.

1977; Tweedie and Spencer 1981).[12] Conversely, export entry and domestic conditions might be positively related if lagged GDP growth reflects growing conditions, say, for agricultural exports. Once a firm has decided to export, factors such as foreign market size will determine the relative attractiveness of each potential location.

3 Data

3.1 Longitudinal Business Database

This paper uses the prototype Longitudinal Business Database (LBD) developed by Statistics New Zealand.[13] This database contains longitudinal administrative and survey data on all "economically significant" firms in the New Zealand economy.[14] From the LBD we use the Longitudinal Business Frame (LBF), which provides information on industry, location and ownership; administrative data from the Inland Revenue Department including goods and services tax (GST) returns, financial accounts (IR10), and company tax returns (IR4); information on employers, employees and wages aggregated to the firm level from the Linked Employer-Employee Dataset (LEED); shipment level merchandise trade data provided by the New Zealand Customs Service (Customs);[15] and value-added data from the Annual Enterprise Survey (AES).

The LBD is predominantly an enterprise level dataset. While each enterprise represents a distinct legal unit, not all enterprises operate independently from others. In particular, groups of firms with parent-subsidiary linkages may operate in a vertically-integrated manner, with the products of the manufacturing firm being recorded as exports by a linked firm further up the production chain. To allocate recorded export activities back to the production unit we use the export-allocation algorithm developed in Fabling and Sanderson (2010). We use the term *firm* to mean both independent

[12] Note that these studies were completed before the economic reforms of the 1980s and the relationships may have changed dramatically since that time. To our knowledge, this type of analysis has not been performed for New Zealand since the reforms.

[13] Detailed information about the LBD is provided in Fabling (2009).

[14] The threshold for economic significance is an annual turnover of 30-40 thousand NZ dollars, being the point at which firms must file a Goods and Services Tax return.

[15] Adjustments have been made to the classification system to maintain a constant definition of both products and countries over time, as discussed in Fabling and Sanderson (2010).

enterprises and the small number of related manufacturers grouped together using this algorithm.

Examination of past export experience is also complicated by breaks in longitudinal enterprise identification numbers. We mitigate the potential for this issue to affect our measures of export experience by considering only those enterprises (or groups of enterprises) which were active in each of the years from 2000 to 2006 (the period over which all data sources are available).[16]

Finally, because we focus on *subsequent* export market choices we exclude firms with no observed exports over the period 1996-2006 (the period over which we have consistently linked export data) and include firms in the analysis only after their initial entry into exporting. This will tend to bias our population towards high-performing firms, as these firms will be more likely both to survive throughout the period and to have observed exports. By compressing the distribution of firm performance to the higher end this may in turn alter the estimated impact of performance on export entry.[17]

We consider the export performance of each firm over five financial years, 2002-2006, with quarterly observations of their export activities.[18] The final population includes 3,483 manufacturing firms, with between 2,286 and 2,919 firms included in each year,[19] and captures 71.8 percent of aggregate merchandise trade over the period 2002-2006.

3.2 Explanatory variables

Given the strong empirical relationship between firm performance and first time export entry, we include two lagged firm performance variables – log of employment (lag_ln_emp) and multi-factor productivity relative to the industry-year average (lag_mfp). We also include a dummy variable distinguishing independent enterprises from groups of linked manufacturers ($multi\_ent$).

[16] This fix is imperfect, as enterprise number breaks which occur between 1996 and 1999 cannot be observed. Thus, export histories will be partial for a small number of firms.

[17] Fabling and Sanderson (2009) show that firm size (employment), capital intensity and multi-factor productivity are significant determinants of a firm's propensity to commence exporting and to enter into new geographic markets, but are also important in predicting whether the firm will have missing performance data in later years (a proxy for firm closure).

[18] The estimation period is constrained by the need for lagged employment data.

[19] All firm counts in this paper have been random rounded to base three in accordance with Statistics New Zealand confidentiality requirements.

We explicitly allow for export experience to determine entry decisions by including indicators of firms' past trade history. A range of studies have shown the importance of past behaviour in determining current export activity (eg, Bernard and Wagner (2001), Greenaway and Kneller (2004), Campa (2004)). In examining each potential new relationship, we look at whether the firm has previously exported other goods to the same country, or the same product to other countries. We allow for experience to depreciate over time by defining these variables as the inverse length of time since a firm last dealt with that product or country. Thus, the variables $firm\_hist$, *prod_hist* and *cty_hist* will be equal to: zero if the firm has no experience at all, in that product or country; one if they exported in the most recent period; and somewhere between zero and one if they exported in a prior period, depending on the vintage of the most recent experience.[20] We also allow for less direct experience to have an influence on later behaviour by including measures of experience in exporting "similar" products to the country in question – other products in the same HS4 category (*sim_hist_prod*) – or exporting the relevant product to "similar" countries – either geographically close to each other (*sim_hist_region* or *sim_hist_contig* for countries in the same geographic region or sharing a land border), or sharing a common language other than English (*sim_hist_lang*). Again, these variables are expressed in terms of the inverse time since the firm's most recent experience in a relevant relationship.

Clearly exporting is not the only way in which firms may learn about other potential markets. Other forms of engagement such as FDI, joint ventures, offshore production and direct imports also build firms' knowledge of, and experience dealing with, international markets. Our dataset provides some indications of these alternative forms of international engagement (though not a comprehensive set of measures). We include an indicator of foreign ownership (*nr_control*) and a full set of import history variables: $firm\_hist\_m$, *prod_hist_m*, *cty_hist_m*, and *reln_hist_m* where each is defined as the inverse number of quarters since the firm last imported (at all, this product, from this country, or this product from this country).

In defining demonstration variables, we assume that firms will learn best through direct observation of firms in the same local area. We include two sets of demonstration variables, one reflecting employment in incumbent exporters and the other employment in entering exporters. While the activities of incumbent exporters are likely to be more visible and may pro-

[20] Recall, however, that we constrain the population to firms which have some past export experience. That is, we exclude firm-quarter observations where $firm\_hist = 0$.

vide a better example to follow (given that the incumbents have presumably had some success in maintaining their exports over time), newly entering exporters may provide more information about changing conditions in the relevant product and geographic markets. Our demonstration variables are then the proportion of employment in manufacturing firms in the same regional council[21] which have continued or commenced (in the past twelve months): exporting (*demo_incumbents*, *demo_entrants*); exporting to the country in question (*demo_cty_incumbents*, *demo_cty_entrants*); exporting a similar product (*demo_prod_incumbents*, *demo_prod_entrants*); or both (*demo_reln_incumbents*, *demo_reln_entrants*).

To reflect the likely benefits of targeting large, rich, open and growing economies we include annual estimates of population, GDP per capita and import intensity in destination markets (*ln_pop*, *ln_gdp_pc*, *ln_imp_intensity*) and their three year growth rates (*d3_ln_pop*, *d3_ln_gdp_pc*, *d3_ln_imp_intensity*).

Monthly bilateral exchange rate measures are used to indicate the purchasing power of foreign buyers. In all cases, the exchange rate is defined as foreign currency units per New Zealand dollar and the measures used are deviations of bilateral exchange rates from their 36 month rolling average. Thus, values above (below) one imply that the New Zealand dollar is above (below) its historical mean. As a high New Zealand dollar is expected to dampen trade, we would expect to see an increase in the exchange rate also dampening export market entry. We estimate our models using both nominal (e) and real (r) exchange rates.

We include the annual change in New Zealand GDP (*d1_ln_NZGDP*) as an indicator of domestic demand conditions. Finally, we include distance from New Zealand (*ln_dist*) to capture the effect of physical distance on both the fixed and marginal costs of exporting. A full list of explanatory variables is provided in appendix A, which also provides detail on the source and construction of each variable.

3.3 Actual and potential entry events

We define a relationship entry as being the first time a firm is observed to export a given product to a given country since January 1996 (the earliest consistently available firm-level export data). As all firms in our population have export experience, each entry event involves either the addition

[21] New Zealand is divided into 16 regional councils, with populations ranging from 1.4 million in Auckland to 32,000 in the West Coast (Statistics New Zealand 2008).

of a new product or country to the firm's existing export portfolio, or a new combination of existing export products and countries. In each quarter, and for each product-country combination, a firm can be either an entrant ($reln\_entry = 1$), a potential exporter ($reln\_entry = 0$), or an incumbent exporter. Incumbents, including firms which have exported the relevant product-country combination in the past, are excluded from the estimation, as they do not have the potential to enter that relationship for the first time.

To clarify these possibilities, consider a world of three possible export destinations – Australia, Tonga and Niue – and a (hypothetical) exporting firm – NZ Toasters Ltd. At time $t - n$, the firm is observed to export toasters to Tonga. In period t, they commence exporting toasters to Niue as well. Thus, at time t toasters to Niue are a new relationship for the firm ($reln\_entry = 1$), toasters to Australia remain a potential, but not actual, relationship ($reln\_entry = 0$), and toasters to Tonga are an incumbent relationship (excluded from the analysis of entry).

In order to estimate our variant of equation 1, we need to define the full set of firm-country-product relationships which have the potential to exist. Defining potential entries is complicated. In principle, all firms have the potential to export any good to any country. As our data covers some 13,300 products, 224 destinations and 3,483 active firms this implies there are around 10.4 billion possible trade relationships. With 20 quarters of data we could have as many as 208 billion observations of non-entry. In reality, however, no firm could reasonably be expected to export every possible product. We therefore take a number of steps to limit the definition of potential entry.

Firstly, we restrict the number of products a firm could possibly produce. We assume that for every product exported by firms in a given three-digit ANZSIC manufacturing industry, that product is a potential export for all other firms in the same three-digit industry.[22] That is, if some firms in the Electrical Equipment and Appliance Manufacturing industry export toasters, then every other firm in that industry has the potential to export toasters.[23]

[22] ANZSIC is the Australian and New Zealand Standard Industry Classification, 1996. There are 46 three-digit ANZSIC manufacturing industries.

[23] Multi-enterprise firms are excluded from the definition of industry exports because it is not generally possible to associate these firms with a single manufacturing industry. Some single-enterprise firms export products which no other firm in their industry exports, which they export on only a small number of occasions, and which do not appear to be sensible products for their industry. One-off sales of capital equipment probably explain some of these events. We restrict our definition of potential products to those for which there are at least two firms in the industry exporting within the same four-digit HS category.

Table 1
Proportion of export value and entry events captured

Population definition and restrictions	Trade	Entries
Total aggregate merchandise exports	1.000	
Exports allocated to manufacturing firms	0.769	
Firm has positive employment in all seven years	0.718	1.000
Firm has some past export experience	0.718	0.987
Product exported by ≥ 1 independent manufacturer	0.684	0.942
Similar product exported by ≥ 2 firms in the industry	0.667	0.892
Firm has complete performance data available in year	0.652	0.768
Full set of macro-economic variables available		
Including nominal exchange rates	0.627	0.713
Including real exchange rates	0.618	0.696

Our choice of macro-economic variables restricts the sample to 191 countries for which we have monthly nominal exchange rates as well as annual GDP, population and import intensity. This is reduced to 153 countries when using real exchange rates.[24]

The combination of these restrictions means our population covers 61.8 percent of aggregate trade (table 1). Over the period 2002-2006 we observe a total of 82,983 actual relationship entry events and some thirteen billion observations of potential entry. In keeping with the findings of Fabling and Sanderson (2010) for aggregate export value, the vast majority of actual relationship entries build on existing experience, with firms exporting either new or existing product lines to countries they had already exported to in the past (table 2).

While entry appears to be a very rare event when viewed from the perspective of the range of possible entries which could occur, from the firm's perspective it is much less unusual. In any given quarter around one third of firms enter at least one relationship. Of those firms, around 40 percent enter a single new relationship and a further 40 percent enter less than six new relationships (figure 1). The distribution of entry events has a long tail with around one percent of firms entering more than 25 new relationships in a quarter.

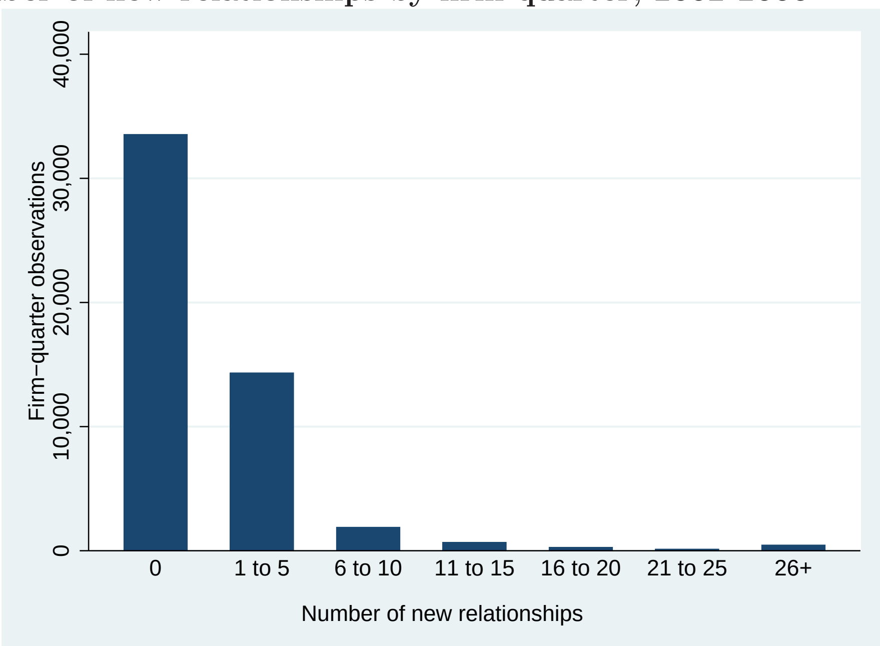
[24] A full list of the countries covered is included in appendix B.

Table 2
Number of entry events by type

New product, old country	34,824
Old product, new country	6,373
New product, new country	2,800
New combination of existing	38,986
Total	82,983

A "new combination of existing" involves firms sending a product from their existing product range to a country they already export to. This is an entry because the mix of product and country has not been observed before.

Figure 1
Number of new relationships by firm-quarter, 2002-2006



4 Methodology

Thirteen billion observations remains an infeasibly large population over which to estimate an empirical model. Further, with only 82,983 observations of actual entry events in the population (79,040 restricting to countries with real exchange rate data), the estimator must be appropriate for rare event models. To address the population size issue, we adopt a case-control sampling strategy,[25] estimating over the entire population of actual entries and a random sample of potential entries to make up a total sample size of two (1.5) million observations using nominal (real) exchange rate data.[26]

We implement the prior correction method for case-control studies of rare events described by King and Zeng (2001, 2004), utilising the ReLogit package in Stata.[27] This approach corrects for selection on the dependent variable while also taking account of uncertainty in the underlying population size. While we can calculate the exact number of entries and non-entries in the population, changing the potential products definition could yield substantially different population sizes. We therefore apply a reasonably wide band around the observed proportion of entry events.[28] However, if the population definition is seriously flawed, mis-estimating the entry rate is probably not the biggest issue, as we may also have bias in our pool of potential entrants. This possibility is addressed in section 6 by considering a substantially more restrictive definition of potential export products.

Table 3 presents population statistics for explanatory variables using the two million observation (nominal exchange rate) sample, weighted to reflect the underlying population distributions.[29] Distributions for some variables, in particular those associated with product- and country-specific trade histories,

[25] Also known as choice-based or endogenous stratified sampling.

[26] King and Zeng (2001) discuss criteria for determining the appropriate number of non-events to include in a rare event model. They suggest that two to five times as many non-events as events is sufficient, as the marginal information provided by each non-event falls as the number of non-events exceeds the number of events. As our sample is limited only by computing power (rather than, say, data collection costs), we use a substantially higher proportion.

[27] See appendix C for a summary of this methodology and the motivation for using it. We also compare alternative methods in section 6.

[28] The proportion of actual entries in the pool of potential entry events is around 6.45×10^{-06} for the sample of countries with nominal exchange rates, and 7.67×10^{-06} for those countries with real exchange rates. In estimation we set the bounds as 10^{-6} and 10^{-5} for both samples.

[29] All results are based on the weighted sample. Regression models also include a full set of quarter and two-digit ANZSIC industry dummies.

are extremely skewed with less than five percent of potential firm-country-product observations having any past experience with the country or product in question. This skewness is an artifact of the definition of potential entry. That is, while the median New Zealand exporter exports only three products to two countries (Fabling and Sanderson 2010), we allow for firms to export to up to 191 countries and between 17 and 2,485 products, depending on their industry. Thus, the chances of a firm having exported a specific good or to a specific country are slim.

5 Results

Results of the empirical estimation are presented as relative risks in table 4. *Relative risk* (or the *risk ratio*) is defined here as $(P|_{x=b})/(P|_{x=a})$ for changes in the explanatory variable x (from a to b), holding all other variables at their mean. We focus on confidence intervals, rather than point estimates, since the former allow for uncertainty in the underlying population incidence rate. For binary variables, the risk is calculated for a 0 to 1 change. Where possible, relative risks for continuous variables are calculated as transitions from the 25th to the 75th percentile. Where the 25th and 75th percentile values are identical, we use the 5th and 95th percentiles or, in cases where the variable is almost always zero (eg, similar history variables), a 0 to 1 change.

Some variables are intrinsically linked together, eg, a firm cannot export a certain product in the previous period (*prod_hist=1*) without also exporting in that period (*firm_hist=1*). Where the values set for the variable of interest bind the values of other variables, we report first the impact of the relevant change in the bound variable (eg, the change from mean to 1 in $firm\_hist$) then the combined effect of that and the specific change we are focusing on (eg, the combined effect of the change from mean to 1 in $firm\_hist$ with a change from 0 to 1 in *prod_hist*).

In interpreting the risk ratios for firm variables, including own export experience, it is important to keep in mind the population definition. The estimated effects are conditional on the firm employing for seven years and having some past export experience. That is, we do not test whether larger, more productive firms are more likely to enter exporting, but rather whether larger, more productive exporters are more likely to expand the range of products and countries in their export portfolio. The results in the top section of table 4 suggest that larger firms, those under domestic ownership and control, and single enterprise firms show a stronger probability of entry into

Table 3
Descriptive statistics

Variable	mean	sd	p5	p25	p75	p95
reln_entry	6.45E-06	2.54E-03	0	0	0	0
Firm characteristics						
lag_ln_emp	2.689	1.539	0.560	1.642	3.558	5.426
lag_mfp	0.112	0.608	-0.791	-0.169	0.427	1.006
nr_control	0.124	0.330	0	0	0	1
multi_ent	0.063	0.243	0	0	0	1
Macroeconomic conditions						
ln_pop	15.200	2.398	10.673	13.574	16.819	18.664
ln_gdp_pc	7.781	1.562	5.442	6.642	8.966	10.352
ln_imp_intensity	-0.718	0.711	-1.879	-1.101	-0.319	0.213
ln_dist	9.464	0.396	8.503	9.352	9.711	9.850
d3_ln_pop	0.000	0.022	-0.002	-0.000	0.000	0.002
d3_ln_gdp_pc	0.001	0.090	-0.101	-0.024	0.026	0.105
d3_ln_imp_intensity	0.003	0.329	-0.387	-0.093	0.104	0.398
d1_ln_NZGDP	0.032	0.012	0.016	0.027	0.040	0.049
Demonstration effects						
demo_incumbents	0.685	0.045	0.583	0.667	0.716	0.725
demo_prod_incumbents	0.088	0.086	0.000	0.018	0.140	0.259
demo_cty_incumbents	0.092	0.113	0	0.001	0.131	0.332
demo_reln_incumbents	0.001	0.008	0	0	0	0.002
demo_entries	0.017	0.009	0.007	0.013	0.018	0.032
demo_prod_entries	0.011	0.017	0	0.001	0.014	0.042
demo_cty_entries	0.007	0.014	0	0	0.007	0.032
demo_reln_entries	0.000	0.004	0	0	0	0.000
Exchange rates						
e	1.158	0.385	0.905	1.032	1.224	1.466
e_high	0.987	0.596	0	1.032	1.224	1.466
Own trade experience						
firm_hist	0.655	0.402	0.053	0.200	1	1
prod_hist	0.005	0.060	0	0	0	0
cty_hist	0.017	0.117	0	0	0	0
sim_hist_prod	0.000	0.017	0	0	0	0
sim_hist_region	0.000	0.017	0	0	0	0
sim_hist_lang	0.000	0.010	0	0	0	0
sim_hist_contig	0.000	0.007	0	0	0	0
firm_hist_m	0.662	0.423	0	0.167	1	1
prod_hist_m	0.010	0.084	0	0	0	0
cty_hist_m	0.018	0.121	0	0	0	0
reln_hist_m	0.000	0.007	0	0	0	0

Calculated from 2,000,000 observation sample, weighted to reflect the original population. Statistics reported as 0.000 are not precisely zero. Variables defined in appendix A.

Table 4
Relative risks

		Point	95% CI	
Estimated probability of entry at mean values of explanatory variables		3.18E-07	5.07E-08	5.85E-07
Firm characteristics				
lag_ln_emp	p25 → p75	1.372	1.243	1.500
lag_mfp	p25 → p75	1.009	0.953	1.064
nr_control	0 → 1	0.840	0.713	0.967
multi_ent	0 → 1	0.535	0.428	0.641
Macroeconomic conditions				
ln_pop	p25 → p75	1.027	0.934	1.120
ln_gdp_pc	p25 → p75	1.163	1.040	1.285
ln_imp_intensity	p25 → p75	1.173	1.131	1.216
ln_dist	p25 → p75	0.905	0.869	0.941
d3_ln_pop	p25 → p75	1.000	1.000	1.000
d3_ln_gdp_pc	p25 → p75	1.005	0.989	1.022
d3_ln_imp_intensity	p25 → p75	0.995	0.973	1.016
d1_ln_NZGDP	p25 → p75	0.988	0.951	1.025
Demonstration effects				
demo_incumbents	p25 → p75	0.976	0.895	1.058
demo_prod_incumbents	p25 → p75	1.395	1.344	1.447
demo_cty_incumbents	p25 → p75	2.159	1.993	2.325
demo_reln_incumbents	p5 → p95	1.002	1.000	1.003
demo_entries	p25 → p75	1.024	1.005	1.043
demo_prod_entries	p25 → p75	1.020	1.006	1.035
demo_cty_entries	p25 → p75	1.089	1.079	1.098
demo_reln_entries	p5 → p95	1.001	1.000	1.001
Exchange rates				
e	p5 → 1	0.939	0.918	0.961
e & e_high	1 → p75	0.894	0.858	0.931
Own trade experience				
firm_hist	p25 → p75	1.923	1.719	2.128
firm_hist	mean → 1	1.323	1.262	1.384
& prod_hist	0 → 1	85.131	74.106	96.156
& cty_hist	0 → 1	28.141	24.784	31.498
firm_hist & cty_hist	mean → 1	26.485	23.240	29.730
& sim_hist_prod	0 → 1	1500	980	2115
firm_hist & prod_hist	mean → 1	82.869	72.215	93.523
& sim_hist_region	0 → 1	1200	850	1585
& sim_hist_lang	0 → 1	210	130	277
& sim_hist_contig	0 → 1	940	410	1475
firm_hist_m	p25 → p75	1.608	1.433	1.783
firm_hist_m	mean → 1	1.210	1.158	1.262
& prod_hist_m	0 → 1	9.422	8.059	10.784
& cty_hist_m	0 → 1	1.654	1.423	1.884
firm_hist_m & prod_hist_m & cty_hist_m	mean → 1	12.364	9.959	14.769
& reln_hist_m	0 → 1	770	350	1185

Rare events logit model using prior correction method to account for case-control sampling. Estimated in Stata9 using ReLogit package (Tomz et al 1999). $\tau \in [0.000001, 0.00001]$. Variables defined in appendix A.

new export relationships. Firms at the 75th percentile in terms of their employment have between 24 and 50 percent higher probability of relationship entry than those at the 25th percentile, while the relative probabilities of foreign-owned and multi-enterprise firms are 3 to 29 and 36 to 57 percent lower respectively, when all other variables are held at their means.[30]

The negative estimated effect of being a multi-enterprise firm may be in part an artifact of the industry-based definition of potential products. That is, we do not include multi-enterprise firms in the definition of potential products by industry (thus missing some products which are exported only by these enterprise groups and reducing the number of actual events we see for these groups) while at the same time we allow for them to export the products associated with the industries of all their constituent manufacturing enterprises (thus increasing the number of non-events).

Meanwhile, the lower entry probability for foreign-owned firms may suggest that market-seeking (rather than resource-, efficiency- or asset-seeking) is the dominant motivation for their establishment in New Zealand, or that the exports of foreign-owned firms are more limited in the range of products or countries involved (eg, exporting only to the country of the parent firm).[31]

Keeping the reference group clearly in mind is also important with respect to own export experience variables. As we consider firms only in quarters after their first observed export activity, $firm\_hist$ (the inverse of the number of quarters since the firm last exported) is constrained to be greater than zero, while all other experience variables can be – and in most cases actually are – zero. The results for own experience variables suggest that closely related forms of export experience (such as sim_hist_prod and sim_hist_region)[32] dramatically increase the probability of additional relationship entry. Meanwhile the effect of $firm\_hist$ shows that very recent export experience (in the previous quarter) is associated with between 72 and 113 percent higher chance of entry into a new relationship relative to a firm which last exported five quarters previously.

[30] Employment may proxy for other correlated characteristics of the firm. For example, when the diversity of export experience (log of the total number of past trade relationships the firm has been involved in) is included in the model as an alternate measure of firm size, the relationship between employment and subsequent relationship entries becomes insignificant but other variables are not strongly affected.

[31] Manova and Zhang (2009) find that although foreign affiliated and joint venture firms in China trade more and exhibit more diversified imports, they export fewer products to fewer destinations than private domestic firms.

[32] Experience exporting a similar product to the country in question, or the same product to another country in the same region, respectively.

Firms with import experience in a specific country show between 40 and 90 percent higher chances of entering a new relationship with that country. In contrast, firms with experience importing a specific product show an eight to eleven-fold higher probability of entering a new relationship involving that good. Greater emphasis on the product dimension may reflect the "product cycle" model of Vernon (1966), in which importers of a product subsequently learn to produce and eventually export the product, and/or a mixed production and distribution model where diversified producer-distributors capitalise on economies of scale and scope in their domestic distribution systems by importing foreign varieties and marketing them domestically while simultaneously producing and exporting their own varieties.[33]

Richer and more open countries are more common targets for new relationships (*ln_gdp_pc* and *ln_imp_intensity*). However, the impact of destination macroeconomic characteristics shows up only with respect to the levels differences, rather than differences in growth rates. Similarly, changes in domestic conditions in New Zealand show no significant association with relationship entry.

Countries that are closer to New Zealand are also more common targets. The difference between the 25th and 75th percentiles of *ln_dist* (approximately 5,000 kilometres) is associated with between 6 and 13 percent decrease in the probability of relationship entry.

The results also suggest that appreciations of the New Zealand dollar have a negative effect on relationship entry. We allow for a different slope to the relationship depending on whether the exchange rate is above or below its 36 month historical average (whether e is above or below 1) by including both the exchange rate variable itself (e) and the same variable interacted with a dummy equal to one if $e > 1$.[34]

We calculate two sets of risk ratios for the exchange rate variable. In table 4, we consider the impact of a change from the 5th percentile to 1 (parity with the 36 month historical average), and that of a change from 1 to the 75th percentile. However, as the magnitude of the latter change is much larger (a difference of 0.224 rather than 0.095) we also compare over equal distances

[33] Alternatively firms may import goods, make minor alterations or repairs, and re-export them under the same product classification. In an (unreported) robustness test we allow the import history coefficients to differ for firms that have re-exported previously. The estimated impact of product imports on subsequent entry is lower for re-exporting firms, suggesting that re-exporting does not explain the stronger product effect.

[34] The New Zealand dollar went through a period of appreciation before and during the estimation period, meaning that $e > 1$ for over three quarters of observations.

Table 5
Relative risks calculated across equal magnitude changes

		Point	95% CI	
Exchange rates				
e	0.905 → 1	0.938	0.918	0.959
e & e_high	1 → 1.095	0.953	0.937	0.969
Demonstration effects				
demo_incumbents	0.001 → 0.03	0.983	0.935	1.031
demo_prod_incumbents	0.001 → 0.03	1.083	1.073	1.092
demo_cty_incumbents	0.001 → 0.03	1.188	1.166	1.209
demo_entries	0.001 → 0.03	1.147	1.027	1.266
demo_prod_entries	0.001 → 0.03	1.047	1.012	1.081
demo_cty_entries	0.001 → 0.03	1.417	1.366	1.468

See table 4 for notes on estimation method. Relationship-level demonstration effects excluded as the 0.001 → 0.03 transition exceeds the range observed in the data. Variables defined in appendix A.

above and below parity. These results are reported in the top section of table 5, and are calculated over a change of 0.095 either side of 1. While the point estimates suggest a slightly stronger effect at the lower end (6.2 percent rather than 4.7 percent), and t-tests on the underlying logit coefficients show that the difference in slopes is statistically significant (unreported), there is substantial overlap between the confidence bands, implying that the difference is not material. Overall, however, exchange rate movements are important with both relative risks significantly different from one.

In table 6 we compare the exchange rate estimates for nominal and real exchange rates over the sample of 153 countries for which both are available. For comparability, we report relative risks calculated over the same magnitude of change above and below par for both the nominal and real exchange rates. In all cases, the incentive effect of a depreciation in the New Zealand dollar below its historical average appears to be slightly stronger than the disincentive effect of an appreciation. Again, there is substantial overlap between the confidence bands for the relative risks, implying that the effect is not materially different above and below "par". The similarity of the results across the nominal and real exchange rate, and between the nominal results for the larger and smaller country samples (tables 5 and 6) gives us confidence that using the nominal exchange rate (ie, maximising the country coverage) in our main estimates is acceptable.

Table 6
Relative risks for nominal vs real exchange rates

		Point	95% CI	
Nominal exchange rate				
Estimated probability of entry at mean values of explanatory variables		3.20E-07	5.08E-08	5.90E-07
e	0.905 → 1	0.941	0.911	0.970
e & e_high	1 → 1.095	0.954	0.931	0.977
Real exchange rate				
Estimated probability of entry at mean values of explanatory variables		3.18E-07	5.10E-08	5.85E-07
r	0.905 → 1	0.926	0.893	0.959
r & r_high	1 → 1.095	0.949	0.921	0.977

See table 4 for notes on estimation method. Calculated for a sample of 1,500,000 observations across 153 countries for which real exchange rate data was available. For this subpopulation the 5th (75th) percentile of e is 0.910 (1.218) and of r is 0.875 (1.168) respectively. Variables defined in appendix A.

Turning to the evidence for demonstration effects, the results suggest that there are few or no spillovers associated with the general export propensity of firms in the region (the risk ratio for *demo_incumbents* is not significantly different from one). However, there is a tendency for firms to follow in the footsteps of existing exporters in terms of both the products they export and the countries they export to (table 4). This effect appears strongest in relation to the activities of incumbent exporters, rather than new entrants, though this may in part reflect the distribution of the underlying demonstration variables which are more highly skewed towards zero for entry than for incumbent exporters (table 3).

The effect of demonstration variables are thus best understood by considering differences across geographic regions, rather than considering the marginal impact of each additional exporting firm within a region. In particular, firms which are located in New Zealand regions with high shares of employment in incumbent exporters to a specific country will have a probability of entering a new relationship involving that country that is 116 percent higher than those in regions with low incumbent employment shares. The same comparison for product-specific demonstration effects is associated with a 40 percent higher entry propensity. The estimated effect of differences in the share of employment in entering exporters are an order of magnitude lower, ranging from 2 to 9 percent.[35]

If we normalise the changes in the share of regional employment associated with each activity (table 5), we see a somewhat different pattern. Comparing like with like, the relationship is still substantially stronger with respect to countries than products (for incumbents, 7 to 9 percent for products and 17 to 21 percent for countries). However, the share of employment in firms which enter a country for the first time is associated with a stronger demonstration effect than the share of employment in firms which are incumbent in a given country (in contrast to table 4 where *demo_cty_incumbents* had a higher risk ratio). The *demo_cty_entries* variable is more likely to capture changes in conditions in the destination country (eg, changes in trade policy or the availability of transport which make certain countries more attractive but are not observed in our macro variables) than the incumbent measure because of hysteresis in export behaviour. That is, incumbent traders are less likely to convey a clear picture of current export conditions since they have previously "locked in" their export behaviour.

Finally, we note that while several of our explanatory variables are associated with large changes in the relative probability of entry, the overall probability that a potential entry event will be realised remains very low. At the mean value of all the explanatory variables, only one in every 1.7 million potential entries is predicted to be an actual entry (top row of table 4). Thus, even for firms which have exported a similar product to the very same country the previous quarter ($sim\_hist\_prod = 1$), there is only a one in 1,000 chance they will commence a new relationship with that country in the following quarter. This is not surprising, given the broad definition of potential entry events, in which many firms are potential exporters of over 1,000 products and have the potential to export to 191 countries.

[35] Alternative measures of demonstration effects according to the share of firms in the region which were observed in the relevant export activity were also considered. The employment based definition is preferred because it weights each firm according to its relative "visibility" (the activities of large firms are more likely to be noticed or the chances of an employee from one firm interacting with an employee from another firm is much higher when those firms are relatively large).

6 Robustness tests

In this section we test the robustness of our estimation approach. We first discuss the sensitivity of our results to changes in the estimation method. We then present results for a more conservative definition of potential export products, where a firm can only export products "similar" to those it has already exported.

6.1 Estimation methods

The main estimates (table 4) used the prior correction method accounting for uncertainty in the proportion of events in the population. As a sensitivity test we compare those results with a rare events logit model using the alternative weighting method outlined in King and Zeng (2001) (table 7).[36] Between these two models, there are few differences in terms of the direction and significance of estimated risk ratios. The exceptions are four variables (*nr_control*, *ln_gdp_pc*, *ln_dist*, *demo_prod_entries*) which appear to have a significant effect based on the prior correction model, but which have a relative risk which is not significantly different from one in the weighted correction model. Further, while the two models are similar in terms of the relative effect of the different explanatory variables, the weighted correction model tends to predict less extreme effects for the relative risk associated with the past experience variables. This is balanced by a higher overall estimate of the probability of entry (at the mean of all explanatory variables). Overall, the sensitivity test suggests a need to be somewhat cautious about the magnitude of the own-firm experience effect, but reinforces the finding that own-experience is indeed a significant factor in explaining firms' ongoing entry behaviour.

[36] We also estimated (unreported) standard logit and probit models weighted to reflect the case-control sampling method. Perhaps not surprisingly, given the similarities in methodology, the standard logit model generates almost identical results to the rare events logit model presented in this section (table 7). The estimated marginal effects from the probit model are stylistically similar to those of the standard logit model, though the probit specification tends to suggest stronger marginal effects (ie, more akin to those of the prior correction model).

Table 7
Relative risks using weighted rare events logit approach

Scenario		Point estimate	95% CI	
Estimated probability of entry at mean values of explanatory variables		6.03E-07	6.73E-07	7.54E-07
Firm characteristics				
lag_ln_emp	p25 → p75	1.184	1.040	1.366
lag_mfp	p25 → p75	1.052	0.965	1.148
nr_control	0 → 1	0.961	0.815	1.130
multi_ent	0 → 1	0.628	0.505	0.816
Macroeconomic conditions				
ln_pop	p25 → p75	1.128	0.985	1.299
ln_gdp_pc	p25 → p75	0.965	0.842	1.106
ln_imp_intensity	p25 → p75	1.155	1.061	1.256
ln_dist	p25 → p75	0.962	0.913	1.013
d3_ln_pop	p25 → p75	1.000	1.000	1.000
d3_ln_gdp_pc	p25 → p75	0.965	0.923	1.007
d3_ln_imp_intensity	p25 → p75	1.005	0.960	1.054
d1_ln_NZGDP	p25 → p75	1.018	0.951	1.090
Demonstration effects				
demo_incumbents	p25 → p75	1.010	0.924	1.101
demo_prod_incumbents	p25 → p75	1.356	1.249	1.470
demo_cty_incumbents	p25 → p75	2.026	1.827	2.253
demo_reln_incumbents	p5 → p95	1.000	0.998	1.002
demo_entries	p25 → p75	1.032	1.005	1.061
demo_prod_entries	p25 → p75	1.030	0.998	1.063
demo_cty_entries	p25 → p75	1.076	1.055	1.098
demo_reln_entries	p5 → p95	1.001	1.000	1.002
Exchange rates				
e	p5 → 1	0.939	0.893	0.990
e & e_high	1 → p75	0.912	0.840	0.992
Own trade experience				
firm_hist	p25 → p75	2.907	2.561	3.268
firm_hist	mean → 1	1.581	1.497	1.668
& prod_hist	0 → 1	41.777	31.644	55.317
& cty_hist	0 → 1	22.733	19.385	26.920
firm_hist & cty_hist	mean → 1	21.712	18.563	25.800
& sim_hist_prod	0 → 1	130	100	172
firm_hist & prod_hist	mean → 1	40.579	30.888	54.054
& sim_hist_region	0 → 1	230	160	315
& sim_hist_lang	0 → 1	54.437	27.760	104.248
& sim_hist_contig	0 → 1	54.366	25.780	111.506
firm_hist_m	p25 → p75	1.809	1.556	2.111
firm_hist_m	mean → 1	1.274	1.199	1.359
& prod_hist_m	0 → 1	5.556	4.308	7.240
& cty_hist_m	0 → 1	1.367	1.144	1.614
firm_hist_m & prod_hist_m & cty_hist_m	mean → 1	5.838	4.409	7.755
& reln_hist_m	0 → 1	16.590	7.445	36.750

Rare events logit model using weighting method to account for case-control sampling. Estimated in Stata9 using ReLogit package (Tomz et al 1999). $\tau \in [0.000001, 0.00001]$. Variables defined in appendix A.

6.2 Re-defining potential products

Finally, we examine the impact of altering the definition of potential export products. Specifically, we restrict our population of both actual and potential entry events by requiring that for a certain HS ten-digit good to be a potential export product for a firm, that firm must have exported a good in the same four-digit HS group in the past.[37]

Table 8 reports the relative risk results for this restricted population. The restrictions lead to a substantial reduction in population size - from nearly 13 billion potential entry events to a little over 1.2 billion. This is mainly due to a reduction in the number of products per firm, rather than the number of firms over which we estimate the model. The remaining population is more heavily weighted towards actual entries than the original sample. Around three-quarters of the initial population of 82,983 actual entry events were in firms which had some past export experience in a similar product line. The fact that we lose a quarter of actual entries implies that the narrower definition of potential export products is too tight – the reason we prefer the broader definition of potential entry. In contrast, less than ten percent of the initial population of non-entries involved relationships in which the firm had similar past experience. Thus, the overall probability of entry is substantially higher in this restricted population – around four times higher overall, and almost twelve times higher when evaluated at the mean of the explanatory variables.[38] The revised incidence rate, τ, sits outside the bounds set for the main model. Allowing for uncertainty in the population incidence rate we set $\tau \in [0.00001, 0.0001]$ for the restricted regressions.

Despite the extreme change in the potential product assumption, the estimates in tables 4 and 8 are remarkably similar. The key patterns associated with own-firm export experience remain strong, as do those associated with destination market characteristics. Only two significant variables (lagged employment and the relationship-level demonstration effect from incumbent exporters) change in sign.[39] Among the import history and demonstration variables, the existing patterns are still evident but only the strongest relationships remain significant in the restricted sample.

[37] This remains a fairly broad definition in the case of many product groups – for example, electric water heaters, hairdryers and coffee makers all come under the same four-digit heading.

[38] The means themselves have also changed as those observations involving firm-product-quarters with no past experience have been dropped from the dataset.

[39] For *lag_ln_emp*, the relative risk is barely significantly different from zero.

Table 8
Relative risks using restricted definition of potential products

Scenario		**Point estimate**	**95% CI**	
Estimated probability of entry at mean values of explanatory variables		6.22E-07	3.94E-06	7.26E-06
Firm characteristics				
lag_ln_emp	p25 → p75	0.904	0.810	0.998
lag_mfp	p25 → p75	0.986	0.929	1.043
nr_control	0 → 1	0.774	0.672	0.877
multi_ent	0 → 1	0.637	0.536	0.738
Macroeconomic conditions				
ln_pop	p25 → p75	1.162	1.047	1.278
ln_gdp_pc	p25 → p75	1.322	1.194	1.450
ln_imp_intensity	p25 → p75	1.134	1.093	1.176
ln_dist	p25 → p75	0.841	0.807	0.875
d3_ln_pop	p25 → p75	1.000	1.000	1.000
d3_ln_gdp_pc	p25 → p75	1.009	0.993	1.025
d3_ln_imp_intensity	p25 → p75	0.979	0.958	1.000
d1_ln_NZGDP	p25 → p75	1.044	1.009	1.079
Demonstration effects				
demo_incumbents	p25 → p75	0.971	0.899	1.042
demo_prod_incumbents	p25 → p75	0.993	0.948	1.038
demo_cty_incumbents	p25 → p75	2.053	1.897	2.208
demo_reln_incumbents	p5 → p95	0.978	0.973	0.984
demo_entries	p25 → p75	1.013	0.994	1.031
demo_prod_entries	p25 → p75	0.994	0.974	1.014
demo_cty_entries	p25 → p75	1.092	1.079	1.105
demo_reln_entries	p5 → p95	1.003	0.999	1.006
Exchange rates				
e	p5 → 1	0.972	0.953	0.991
e & e_high	1 → p75	0.942	0.912	0.973
Own trade experience				
firm_hist	p5 → p95	1.313	1.128	1.497
firm_hist	mean → 1	1.035	1.017	1.052
& prod_hist	0 → 1	21.517	19.465	23.569
& cty_hist	0 → 1	18.277	16.202	20.352
firm_hist & cty_hist	mean → 1	15.629	13.960	17.297
& sim_hist_prod	0 → 1	94.757	81.406	108.107
firm_hist & prod_hist	mean → 1	17.918	16.307	19.529
& sim_hist_region	0 → 1	160	140	180
& sim_hist_lang	0 → 1	40.368	32.025	48.711
& sim_hist_contig	0 → 1	87.172	65.525	108.819
firm_hist_m	p5 → p95	1.040	0.894	1.186
firm_hist_m	mean → 1	1.005	0.985	1.024
& prod_hist_m	0 → 1	3.804	3.493	4.116
& cty_hist_m	0 → 1	1.048	0.919	1.177
firm_hist_m & prod_hist_m & cty_hist_m	mean → 1	3.606	3.088	4.123
& reln_hist_m	0 → 1	51.615	35.943	67.288

See table 4 for notes on estimation method. $\tau \in [0.00001, 0.0001]$. Variables defined in appendix A.

7 Conclusion

Overall, the results suggest that sunk costs are a substantial factor determining not only whether firms will expand into new markets, but also which markets and products they will choose when expanding their export relationships. In particular, firms are more likely to introduce additional products to countries with which they already have an established trade relationship. At the same time, the costs of product development imply that firms will also choose to expand by introducing their existing, successful products to new geographic markets. That is, we find strong evidence of path dependence.

There is evidence that product- and relationship-level import experience also play a role in determining the future expansion of export relationships, perhaps driven by some form of "product cycle" or reflecting the operation of diversified producer-distributors.

The results also suggest a role for export propensity spillovers from other domestic firms. These spillovers appear to be relationship-specific, in that a higher general propensity to export in the region has no impact on a firm's probability of entry into new export relationships, yet the observed experience of firms exporting similar products, or exporting to the country in question is associated with a substantial increase in the probability of entry. Both the activities of incumbents and new entrants seem to provide a demonstration effect for potential entrants.

Fabling and Sanderson (2010) show that the expansion of incumbent exporters into new trade relationships accounts for around 60 percent of total growth in aggregate trade in New Zealand between 1996-98 and 2004-06, far outweighing the 12-16 percent contribution of newly entering exporters. As such, even small impacts on the ability of firms to expand their export products and markets may have substantial benefits for aggregate export earnings. In this paper we have shown that the role of past experience has an important impact on firms' future export choices. Path dependence is thus a very real force acting on the overall size and distribution of the aggregate export portfolio.

References

Aitken, B, G Hanson, and A Harrison (1997), "Spillovers, foreign investment, and export behavior," *Journal of International Economics*, 43(1-2), 103–132.

Anderson, J and E van Wincoop (2004), "Trade costs," *Journal of Economic Literature*, 42(3), 691–751.

Baldwin, R and J Harrigan (2007), "Zeros, quality and space: Trade theory and trade evidence," *National Bureau of Economic Research, Working Papers*, 13214.

Bernard, A and J Jensen (2004), "Why some firms export," *The Review of Economics and Statistics*, 86(2), 561–569.

Bernard, A, J B Jensen, S Redding, and P Schott (2009), "The margins of US trade," *American Economic Review*, 99(2), 487–93.

Bernard, A, S Redding, and P Schott (2006), "Multi-product firms and trade liberalization," *National Bureau of Economic Research, Working Paper*, 12782.

Bernard, A and J Wagner (2001), "Export entry and exit by German firms," *Review of World Economics (Weltwirtschaftliches Archiv)*, 137(1), 105–123.

Caliendo, M, F Fossen, and A Kritikos (2009), "Risk attitudes of nascent entrepreneurs – new evidence from an experimentally validated survey," *Small Business Economics*, 32(2), 153–167.

Campa, J (2004), "Exchange rates and trade: How important is hysteresis in trade?" *European Economic Review*, 48(3), 527–548.

Clerides, S, S Lach, and J Tybout (1998), "Is learning by exporting important? Micro-dynamic evidence from Colombia, Mexico, and Morocco," *The Quarterly Journal of Economics*, 113(3), 903–947.

Criscuolo, P (2009), "Inter-firm reverse technology transfer: The home country effect of R&D internationalization," *Industrial and Corporate Change*, 18(5), 869–899.

Das, S, M Roberts, and J Tybout (2007), "Market entry costs, producer heterogeneity, and export dynamics," *Econometrica*, 75(3), 837–873.

Eaton, J, S Kortum, and F Kramarz (2008), "An anatomy of international trade: Evidence from French firms," *National Bureau of Economic Research, Working Paper*, 14610.

Eckel, C and J P Neary (2009), "Multi-product firms and flexible manufacturing in the global economy," *forthcoming in Review of Economic Studies*.

Evenett, S and A Venables (2002), "Export growth in developing countries: Market entry and bilateral trade flows," Mimeo.

Fabling, R (2009), "A rough guide to New Zealand's Longitudinal Business Database," *Institute of Economic Research, Hitotsubashi University, Global COE Hi-Stat Discussion Papers*, No. 103.

Fabling, R and L Sanderson (2009), "Exporting and performance: The impact of destination characteristics on learning effects," paper presented at the 9th Comparative Analysis of Enterprise Data Conference, Tokyo, Japan.

Fabling, R and L Sanderson (2010), "Entrepreneurship and aggregate merchandise trade growth in New Zealand," *Journal of International Entrepreneurship*, forthcoming.

Greenaway, D and R Kneller (2004), "Exporting and productivity in the United Kingdom," *Oxford Review of Economic Policy*, 20(3), 358–371.

Hausmann, R and D Rodrik (2003), "Economic development as self-discovery," *Journal of Development Economics*, 72(2), 603–633.

King, G and L Zeng (2001), "Logistic regression in rare events data," *Political Analysis*, 9(2), 137–163.

King, G and L Zeng (2004), "Inference in case-control studies," in *Encyclopedia of Biopharmaceutical Statistics: Second Edition*, ed S-C Chow, 1–13, Marcel Dekker: New York, 2 edn.

Manova, K and Z Zhang (2009), "China's exporters and importers: Firms, products and trade partners," *National Bureau of Economic Research, Working Paper*, 15249.

Manski, C (1999), "Nonlinear statistical inference: Essays in honor of Takeshi Amemiya," in *Nonparametric Identification Under Response-Based Sampling*, eds C Hsaio, K Morimune, and J Powell, Cambridge University Press: Cambridge.

Manski, C and S Lerman (1977), "The estimation of choice probabilities from choice based samples," *Econometrica*, 45(8), 1977–1988.

McCullagh, P and J Nelder (1989), *Generalized Linear Models*, Chapman and Hall, New York, 2 edn.

Melitz, M and G Ottaviano (2008), "Market size, trade, and productivity," *Review of Economic Studies*, 75(1), 295–316.

Morgan, G (1977), "Modelling exports by New Zealand industries - A preliminary investigation," paper presented at the New Zealand Association of Economists conference, Hamilton, New Zealand.

Rauch, J (1999), "Networks versus markets in international trade," *Journal of International Economics*, 48(1), 7–35.

Statistics New Zealand (2008), "Subnational population estimates: At 30 June 2008," *Statistics New Zealand: Wellington.*

Tomz, M, G King, and L Zeng (1999), "Relogit: Rare events logistic

regression," Statistical software for Stata.
Tweedie, A and G Spencer (1981), "Supply behaviour in New Zealand's export industries," *Reserve Bank of New Zealand*, *Discussion Papers*, G/81/1.
Vernon, R (1966), "International investment and international trade in the product cycle," *The Quarterly Journal of Economics*, 80(2), 190–207.
Wagner, J (2004), "Are young and small firms hothouses for nascent entrepreneurs? Evidence from German micro data," *Institute for the Study of Labor (IZA)*, *Discussion Paper*, 989.
Wagner, J (2007), "Exports and productivity: A survey of the evidence from firm-level data," *The World Economy*, 30(1), 60–82.

Appendices

A Variable definitions

A.1 Data sources

Datasets within the Longitudinal Business Database	
AES	Annual Enterprise Survey
BAI	Business Activity Indicator
Customs	New Zealand Customs Service Import and Export Entry forms
IR4	Company Tax Returns
IR10	Accounts Information
LBF	Longitudinal Business Frame
LEED	Linked Employer-Employee Dataset

External data sources	
IFS	International Financial Statistics database http://www.imfstatistics.org/imf/
UN	United Nations Statistics Division, National Accounts Main Aggregates Database http://unstats.un.org/unsd/snaama/Introduction.asp
IDB	US Census Bureau International Database http://www.census.gov/ipc/www/idb/
CEPII	CEPII Distances dataset http://www.cepii.fr/anglaisgraph/bdd/distances.htm
USDA	US Department of Agriculture http://www.ers.usda.gov/Data/ExchangeRates/

A.2 Variable definitions

	Variable names	Variable definitions	Periodicity	Data sources	Expected sign
Firm characteristics	lag_mfp	multifactor productivity (mfp) in the previous year relative to the industry-year average. Estimated using a Cobb-Douglas production function with non-constant returns to scale	annual	LEED, LBF AES, BAI, IR10	+
	lag_ln_emp	natural log of total employment (employees plus working proprietors) in the previous year	annual	LEED	+
	nr_control	foreign control: dummy = 1 if the manufacturing enterprise is under foreign ownership or control	annual	IR4, LBF	?
	multi_ent	dummy variable identifying multi-enterprise groups	annual	LBF	?
	industry	three-digit ANZSIC96 industry code	annual	LBF	?
Export experience	firm_hist	inverse number of quarters since the firm last exported (any product to any destination)	quarterly	Customs	+
	prod_hist	inverse number of quarters since the firm last exported this product	quarterly	Customs	+
	cty_hist	inverse number of quarters since the firm last exported to this country	quarterly	Customs	+
Import experience	firm_hist_imp	inverse number of quarters since the firm last imported (any product from any source)	quarterly	Customs	+
	prod_hist_m	inverse number of quarters since the firm last imported this product	quarterly	Customs	+
	cty_hist_m	inverse number of quarters since the firm last imported from this country	quarterly	Customs	+
	reln_hist_m	inverse number of quarters since the firm last imported this product from this country	quarterly	Customs	+
Export experience - similar relationships	sim_hist_prod	inverse number of quarters since the firm last exported a similar product (same four-digit HS group) to this country	quarterly	Customs	+
	sim_hist_lang	inverse number of quarters since the firm last exported this product to a country sharing a common language (a language other than English spoken by at least 20 percent of the population of both countries)	quarterly	Customs, CEPII	+
	sim_hist_region	inverse number of quarters since the firm last exported this product to a country in the same geographic region as this country	quarterly	Customs	+
	sim_hist_contig	inverse number of quarters since the firm last exported this product to a country which shares a land-border with this country	quarterly	Customs, CEPII	+

Continued on next page

Continued from previous page

	Variable names	Variable definitions	Periodicity	Data sources	Expected sign
Demonstration effects	demo_incumbents	proportion of manufacturing employment in the same regional council area in firms which have exported (anything to any country) in the past 12 months, excluding firms which exported for the first time	quarterly	Customs, LBF	+
	demo_XXX_incumbents	proportion of manufacturing employment in the same regional council area in firms which have exported (the relevant product, country or both) in the past 12 months, excluding firms which did so for the first time	quarterly	Customs, LBF	+
	demo_entries	proportion of manufacturing employment in the same regional council area in firms which have commenced exporting (anything to any country) in the past 12 months	quarterly	Customs, LBF	+
	demo_XXX_entries	proportion of manufacturing employment in the same regional council area in firms which have commenced exporting (the relevant product, country or both) in the past 12 months	quarterly	Customs, LBF	+
Macroeconomic conditions	ln_pop d3_ln_pop	destination country population (in levels and three-year differences)	annual	IDB	+
	ln_gdp_pc d3_ln_gdp_pc	destination country GDP per capita (in levels and three-year differences)	annual	UN, IDB	+
	ln_imp_intensity d3_ln_imp_int	destination import intensity defined as imports/GDP (in levels and three-year differences)	annual	UN	+
	ln_dist	natural log of the distance (km) between New Zealand and the destination country	annual	CEPII	-
	ln_d1_NZ_gdp	one year growth rate of New Zealand GDP	annual	UN	?
Exchange rates	e	deviation of nominal exchange rate with destination country (foreign currency per NZD) from its 36 month historical rolling average	quarterly	IFS	-
	r	deviation of real exchange rate with destination country (nominal exchange rate × $(CPI_{foreign}/CPI_{NZ})$) from its 36 month historical rolling average CPI data is sourced from IFS. For countries which use a shared currency (eg, the USD) and did not have CPI data available, we attribute the real exchange rate with the principal country using that currency (eg, the USA). For China, we use the real exchange rate calculated by Mathew Shane at the US Department of Agriculture	quarterly	IFS, UN, USDA	-

B Countries included

Countries with both nominal and real exchange rate data		
Albania	Greece	Norway
Algeria	Greenland	Pakistan
Andorra	Grenada	Palau
Angola	Guatemala	Panama
Argentina	Guinea-Bissau	Papua New Guinea
Armenia	Guyana	Paraguay
Aruba	Haiti	Peru
Australia	Honduras	Philippines
Austria	Hong Kong	Poland
Bahamas	Hungary	Portugal
Bangladesh	Iceland	Puerto Rico
Barbados	India	Romania
Belgium	Indonesia/Timor-Leste	Russian Federation
Belize	Iran, Islamic Rep. of	Rwanda
Benin	Ireland	Saint Lucia
Bolivia	Israel	Saudi Arabia
Botswana	Italy	Senegal
Brazil	Jamaica	Seychelles
Bulgaria	Japan	Sierra Leone
Burkina Faso	Jordan	Singapore
Burundi	Kazakhstan	Slovakia
Cambodia	Kenya	Slovenia
Cameroon	Kiribati	Solomon Islands
Canada	Korea, Rep. of	South Africa
Cape Verde	Kuwait	Spain
Chad	Kyrgyzstan	Sri Lanka
Chile	Lao Peoples Dem. Rep.	St Kitts and Nevis
China	Latvia	St Vincent and the Grenadines
Colombia	Lithuania	Sudan
Cook Islands	Luxembourg	Suriname
Costa Rica	Macao	Swaziland
Cote D'Ivoire	Macedonia	Sweden
Croatia	Madagascar	Switzerland
Cyprus	Malawi	Syrian Arab Republic
Czech Republic	Malaysia	Tanzania, United Rep. of
Dem. Rep. of Congo	Mali	Thailand
Denmark	Malta	Togo
Dominica	Marshall Islands	Tonga
Dominican Republic	Mauritius	Trinidad and Tobago
Egypt	Mexico	Tunisia
El Salvador	Micronesia, Federated States of	Turkey
Estonia	Moldova, Rep. of	Turks and Caicos Islands
Ethiopia	Mongolia	Tuvalu
Fiji	Morocco	Uganda
Finland	Mozambique	United Kingdom
France	Nauru	United States
Gabon	Nepal	Uruguay
Gambia	Netherlands	Vanuatu
Georgia	Netherlands Antilles	Venezuela
Germany	Niger	Vietnam
Ghana	Nigeria	Zambia

Countries with only nominal exchange rate data		
Afghanistan	Congo	Namibia
Anguilla	Djibouti	Nicaragua
Antigua and Barbuda	Ecuador	Oman
Azerbaijan	Equatorial Guinea	Qatar
Bahrain	Eritrea	Samoa
Belarus	Guinea	San Marino
Bermuda	Lebanon	Sao Tome and Principe
Bhutan	Lesotho	Tajikistan
Bosnia and Herzegovina	Liberia	Ukraine
Brunei Darussalam	Libyan Arab Jamahiriya	United Arab Emirates
Cayman Islands	Maldives	Yemen
Central African Republic	Montserrat	Zimbabwe
Comoros	Myanmar	

C Technical detail

The core estimation method used in this paper is the rare events logit model specification for situations with limited knowledge of the population incidence rate. The models are estimated using the ReLogit suite of Stata programmes created by Tomz et al (1999) to implement the methods described by King and Zeng (2001, 2004). Rare events models have received limited attention in the economics literature. Among the small number of papers using these techniques are Wagner (2004), Caliendo, Fossen, and Kritikos (2009) and Criscuolo (2009). As these estimation methods may be unfamiliar to some readers, this appendix provides a brief explanation of the details and the motivation for using this methodology.

The current analysis of export market entry presents a number of complications beyond those experienced in standard binary dependent variable analyses. As discussed in the main text, potential entry events vastly outnumber actual entries, and the number of non-entries reaches into the billions. At the same time, there is uncertainty about the true ratio of events to non-events. King and Zeng (2001, 2004) outline a series of adjustments to the standard logit model to correct for rare event bias in a case-control sample design, to allow for the uncertainty in the underlying population incidence rate, and also to provide more readily interpretable results.

Consider a binary dependent variable model, in which the observed dependent variable Y_i is equal to 1 if an entry event occurs, and 0 if it does not

$$Y_i \sim Bernoulli(y_i|\pi_i) = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1-\pi_i. \end{cases}$$

The observed variable Y_i is assumed to be the realisation of an unobserved latent variable Y_i^*[40]

$$Y_i^* \sim Logistic(Y_i^*| - \mathbf{X}_i\boldsymbol{\beta}).$$

Our goal is to estimate the probability of relationship entry π_i as a function of the explanatory variables $\mathbf{X}_i$

$$Pr(Y_i = 1|\boldsymbol{\beta})) = \pi_i = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}.$$

[40] Clearly there are many other possible distributions that could be assumed for the latent variable. We directly consider the implications of assuming a normal distribution (ie, a probit model) in section 6. However, we focus on the logit model as the adjustments developed by King and Zeng (2001) cannot be applied to a probit model.

As we use a case-control sampling method, in which we select all the observed events and a random sample of non-events, the observed proportion of entry events in our sample is purely a sampling decision and bears no relationship to the actual share of entry events in the population. King and Zeng (2001) suggest two methods to correct for this sample design: *prior correction* and *weighting*. The prior correction model relies on the result that the MLE logit estimate $\hat{\boldsymbol{\beta}}_1$ is a consistent estimate of the true $\boldsymbol{\beta}_1$ as the case-control sampling method affects only the intercept term β_0. By correcting the intercept term β_0 to reflect the true population incidence rate τ and the sample incidence rate $\bar{y}$ according to the adjustment $\hat{\beta}_0 - ln[(\frac{1-\tau}{\tau})(\frac{\bar{y}}{1-\bar{y}})]$, prior correction can be used to provide consistent and asymptotically efficient estimates of true population probabilities and risk ratios.

Alternatively, the weighting method (Manski and Lerman 1977) weights the data to compensate for differences between the sample incidence rate $\bar{y}$ and the population incidence rate τ, by calculating the weighted log-likelihood

$$\begin{aligned} lnL_w(\boldsymbol{\beta}|y) &= w_1 \sum_{Y_i=1} ln(\pi_i) + w_0 \sum_{Y_i=0} ln(1-\pi_i) \\ &= -\sum_{i=1}^{n} w_i ln(1 + e^{(1-2y_i)\mathbf{X}_i\boldsymbol{\beta}}) \end{aligned}$$

where the weights are $w_1 = \tau/\bar{y}$ and $w_0 = (1-\tau)/(1-\bar{y})$ and where $w_i = w_1 Y_i + w_0(1-Y_i)$.

King and Zeng (2001) note that weighting is preferable to prior correction when the model is mis-specified, but is asymptotically slightly less efficient. Moreover, standard methods of calculating standard errors and applying corrections for rare events are not appropriate for the weighted model (though King and Zeng (2001) provide an alternate specification which can be used). Most crucially, from our perspective, as the population incidence rate is included within the likelihood estimation, there is no simple way for the weighting method to allow for uncertainty in τ. We therefore favour prior correction as our main estimation method, and present the weighted results only as a robustness check.

A second issue King and Zeng (2001) discuss is that logit models are known to be biased in small samples (eg, McCullagh and Nelder 1989), and this bias carries over to the case of rare events, due to the small number of observed events relative to non-events. Moreover, they show that bias in the coefficients is compounded in the estimation of relevant quantities of interest, such as the absolute and relative risks, by failure to account for uncertainty in the estimated coefficients. These biases imply that both coefficients and associated probabilities will be underestimated in the case of rare events.

King and Zeng (2001) show that bias in the coefficients will be reduced as the sample size n increases, but amplified by the rarity of the event (see footnote 7 and appendices of King and Zeng for the derivation). They go on to suggest bias correction methods, as outlined below. In practice, it seems likely that these two factors will counteract each other in our estimation, as we have a large sample size but very rare events. This assumption is borne out by the data, in that estimates using the weighted rare events correction method show very little difference to a simple weighted logit without the rare event correction. However, we maintain the (more technically correct) rare events finite sample corrections in our main estimates (section 5).

King and Zeng (2001) show that the bias in the coefficient can be estimated using weighted least squares as

$$\text{bias}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\boldsymbol{\xi})$$

where $\mathbf{W} = \text{diag}\{\hat{\pi}_i(1-\hat{\pi}_i)w_i)\}$, $\boldsymbol{\xi}_i = 0.5\mathbf{Q}_{ii}[(1+w_i)\hat{\pi}_i - w_i]$ and $\mathbf{Q}_{ii}$ is the diagonal element of $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'$.

This can be estimated by running a weighted least-squares regression with $\mathbf{X}$ as the "explanatory variables", $\boldsymbol{\xi}$ as the "dependent variable," and $\mathbf{W}$ as the weight, and used to create a bias corrected estimate $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \text{bias}(\hat{\boldsymbol{\beta}})$. As well as correcting the bias on the coefficients, this correction also has the benefit of reducing variance, as $V(\tilde{\boldsymbol{\beta}}) = (\frac{n}{n+k})^2 V(\hat{\boldsymbol{\beta}})$ and $(\frac{n}{n+k})^2 < 1$.

Estimates of the absolute risk (and hence the relative risks) can then be computed by averaging over the uncertainty in $\tilde{\boldsymbol{\beta}}$

$$Pr(Y_i = 1) = \int Pr(Y_i = 1|\boldsymbol{\beta}*)P(\boldsymbol{\beta}*)d\boldsymbol{\beta}*$$

through stochastic simulation, where $\boldsymbol{\beta}*$ is the integration dummy, and to summarise estimation uncertainty $P()$ we take the Bayesian viewpoint and use the posterior density of $\boldsymbol{\beta}$, $N(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, V(\tilde{\boldsymbol{\beta}}))$. This method involves taking a random draw of $\boldsymbol{\beta}$ from $P(\boldsymbol{\beta})$, inserting it into $[1 + e^{-\mathbf{X}_i\boldsymbol{\beta}}]^{-1}$, repeating 1,000 times and then averaging over the simulations to give confidence intervals for the actual $Pr(Y_i = 1)$.[41] Relative risks can then be calculated by inserting two chosen levels of $\mathbf{X}_i$ and computing the ratio of the absolute risks.

Finally, King and Zeng (2004) deal with the issue of uncertainty in the underlying population incidence rate τ. In the discussion above, τ is treated

[41] King and Zeng (2001) also discuss analytic methods for computing the risks.

as a known quantity and used directly in the case-control and rare-event corrections to estimate the relationships of interest at the population level. However, as in our study, there may be substantial uncertainty about the population incidence rate.

Past work in this area has used a variety of extreme assumptions, including the "full information assumption" implied in the discussion above, in which complete knowledge of τ is assumed; Manski's (1999) "ignorance assumption," in which no prior knowledge of τ is assumed; and the "rare disease assumption" used in epidemiology, in which τ is assumed to be approximately zero. King and Zeng (2004) suggest an alternate approach which assumes only that τ can be identified within reasonable bounds $\tau \in [\tau_0, \tau_1]$ – the "available information" assumption.[42] The authors describe this as a "robust Bayesian" approach, in that the choice of an interval for τ is not equivalent to imposing a prior density within those bounds (as per a fully Bayesian model), but effectively narrows the possible priors to the subset for which $\int_{\tau_0}^{\tau_1} P(\tau)d\tau = 1$. This method has the benefit of allowing researchers to be specific about their knowledge of τ, neither under- nor over-stating the degree of confidence they have, but means that the results are limited to a statement of the credible interval for the quantity in question, rather than an exact estimate. This in turn implies that our relative risks and any other quantity of interest must be calculated based on these same bounds eg, the appropriate band for the relative risk is $RR \in [min(RR_{\tau_0}, RR_{\tau_1}), max(RR_{\tau_0}, RR_{\tau_1})]$.

Our estimation method must be able to cope with each of these issues (rare events, case-control sampling, and uncertainty about τ). The main estimates thus follows the following procedure (carried out within the ReLogit program). First, a standard logistic regression is run, estimating the slope vector $\hat{\boldsymbol{\beta}}_1$ (which is consistent in case-control models), and the unadjusted constant $\hat{\beta}_0$ (which is not).[43] τ is assumed to lie within the interval $[10^{-6}, 10^{-5}]$, based on the observed population incidence rate of 6.45×10^{-6}. A thousand simulations of $\boldsymbol{\beta}$ are drawn from the posterior density, $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{V}(\hat{\boldsymbol{\beta}}))$. Half of the estimates for the intercept $\hat{\beta}_0$ are then adjusted with respect to the lower bound on τ ($\tau_x = 10^{-6}$) and the other half are adjusted with respect to the upper bound ($\tau_x = 10^{-5}$), using the correction formula

$$\tilde{\beta}_0 - ln\left[\left(\frac{1-\tau_x}{\tau_x}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

[42] This assumption can be relaxed further, by assuming that $\tau \in [\tau_0, \tau_1]$ with a specified probability, then defining a density function for the tails.

[43] Because of the bias in the intercept and the difficulty interpreting the logit coefficients, we do not report these results in the paper.

to give bounds on the estimates of $\tilde{\boldsymbol{\beta}}$. Confidence intervals for the absolute risk are then constructed by simulation (1,000 reps) using each of the two bounds, τ_0 and τ_1. Absolute risks are estimated directly, with relative risks computed as the ratio of the two absolute risks. Reported point estimates are the median value from these simulations, while the 2.5^{th} and 97.5^{th} percentiles give the 95 percent confidence interval.