

How specialised are specialists? Generalisation properties of entries from the 2008 and 2009 TAC Market Design Competitions.

Edward Robinson¹, Peter McBurney², and Xin Yao¹

¹ Cercia, University of Birmingham, Birmingham, B15 2TT, UK
eyr{xin}@cs.bham.ac.uk

² Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
mcburney@liverpool.ac.uk

Abstract. Unlike the classic Trading Agent competition (TAC), where participants enter trading strategies into a market, the TAC Market Design Competition (CAT) allows participants to create rules for their own double auction market and set fees for traders, which they embody in agents known as *specialists*. Although the generalisation properties of traders when the specialist (i.e., the market mechanism) is fixed have been assessed, generalisation properties of specialists have not. It is unclear whether and how a specialist might (intentionally or unintentionally) favour certain trading strategies. We present an empirical analysis of specialists' generalisation abilities in various trading environments. Our results show that specialists can be sensitive to a number of factors, including the other trading and specialist strategies in the environment.

1 Introduction

The growth of e-commerce has led to increased attention to economic markets from within computer science. Within economics, the discipline of *Mechanism Design* uses methods from mathematical game theory, behavioural economics, and computer simulation to aid in the design and analysis of marketplaces, and of the strategies used by traders within these markets. One common type of market is the *double auction*, in which multiple potential buyers and sellers place (respectively) bids and asks, seeking to engage in purchase transactions over some product or commodity. Double auctions are in use by most of the stock and commodity exchanges around the world, although often with different rules. In a global economy, stock exchanges compete against each other for trading business, and, increasingly, against new online markets not tied to traditional physical exchanges.

In such a competitive environment, the precise rules adopted by a marketplace may have important consequences—in attracting (or not) traders to the market, in rewarding (or not) particular trading strategies, and in facilitating (or not) the matching of shouts (bids and asks), and the execution of trades in the market. Thus, a detailed understanding of the different potential rules for double auction markets and their impacts is important, particularly if we

seek the eventual automation of the design of market mechanisms. However, the mathematical analysis of the double auction is intractable, and so computer or human simulations are currently needed to gain this understanding.

With this in mind, a new research tournament was launched in 2007 to promote research into automated mechanism design: the *Trading Agent Competition Market Design (or CAT) Tournament* [1]. The CAT Tournament comprises a series of artificial parallel markets, designed to mirror the competition between global stock markets. These parallel markets, called *specialists*, are created by entrants to the Tournament, and they compete with one another to attract and retain traders, who are potential buyers and sellers of some abstract commodity. The traders are software agents created and operated by the organisers of the CAT Tournament, in a reversal of the usual Trading Agent Competition structure.

In this paper, we describe simulation analyses undertaken using the CAT Tournament platform JCAT¹ with some of the specialists entered into the 2008 and 2009 Tournaments. Our primary research goal is to better understand the characteristics of the mechanisms used by the specialists, particularly in relation to the contexts in which they trade, in order to design more robust mechanisms. Thus, we seek to generalise specialist capabilities and strengths. Our contribution is to demonstrate that the specialists in the CAT Tournaments are not robust against changes in the trader mix, in the competitor mix and in the scoring period. Changing each of these factors leads to some changes in the tournament ranks and/or the game scores achieved by the specialists. The market mechanisms employed by the specialists may thus be seen not to generalise, and so research will be needed to make these mechanisms more robust.

The paper is organised as follows. Section 2 presents a brief summary of the CAT Tournament, and Section 3 presents our research findings in detail. Section 4 presents our conclusions and proposals for future work.

2 CAT Tournament

The organisation and structure of the TAC Market Design (CAT) Tournament is given in the game documents [1]. Here we briefly mention the most important aspects. A CAT game takes place over a number of simulation *trading-days*, each of which consists of a number of *rounds*. Each round lasts a number of *ticks*, measured in milliseconds. The game uses a client server architecture, with the CAT server controlling the progression of the game. CAT clients are either traders (potential buyers or sellers) or specialists (aka markets). All communication between traders and specialists is via the CAT server.

In the standard CAT installation, four different trader strategies are provided. *Zero Intelligence – Constrained (ZIC)* traders [2] essentially place random bids and asks, within constraints. These traders ignore both market state and the history. *Zero Intelligence Plus (ZIP)* traders [3] are modified versions of *ZIC*

¹ <http://jcat.sourceforge.net/>

traders that seek to remain in profit in competition with other traders, using some market history. *RE* traders [4] use a learning algorithm based on a model of human learning, with the most recent surplus or loss guiding the trader's shouting strategy one step ahead. Finally, *GD* traders [5] use past marketplace history of submissions and transactions to generate beliefs over the likelihood of any particular bid or ask being accepted, which is used to guide shouting strategies. *ZIC* are the least, and *GD* are the most, sophisticated of these four types. In addition, all types of traders in the standard CAT installation use an *n*-armed bandit strategy [6] for selecting which specialist to register with on each new trading day.

Specialists have freedom to set market rules in six broad policy areas: *Charging policy*: What charges and fees are imposed by the specialist on traders? *Quote policy*: What limitations, such as a lower bound for bids, does the specialist impose on shouts by traders? *Shout accepting policy*: When does a specialist accept a shout made by a trader? *Matching policy*: How does the specialist match bids and asks made by traders? *Pricing policy*: What are the transaction prices for matched bid-ask pairs? *Clearing policy*: When does the specialist clear the market and execute transactions between matched bids and asks? Following the 2007 CAT Tournament, [7] undertook a series of simulations to infer the policies of the specialists entered in the game. In addition, the research teams behind two of the 2007 CAT specialists have written about their strategies [8, 9].

In the CAT Tournament, specialists know that each trader is one of the four types, but not the overall proportions of each type. Accordingly, the design of a specialist seeking to win the game cannot be optimised for only a subset of trading strategies. In addition, the scoring metric used by the game is multi-dimensional. Games are scored using an unweighted average of three criteria: the proportion of traders attracted to the specialist each day (market share); the proportion of accepted shouts which are matched (transaction success rate); and the share of profits made by the specialist. As with trader types, this multi-dimensionality creates challenges for the optimal design of specialists, since these criteria may conflict. A game-winning strategy may focus on scoring highly on different criteria at different times in the life-cycle of a game, or against different trader types.

Given this game structure, it is easy to see that some specialists may perform better with traders of a particular type, and/or against competing specialists using particular policies. Because the actual CAT Tournaments are only conducted over a limited number of games (typically, three), the performance of a specialist in the Tournament is not necessarily a good guide to that specialist's general ability, i.e., with other trader mixes, or in competition against different specialists. Niu *et al.* [10], for example, have shown that some of the well-performing 2007 CAT specialists have weaknesses in other situations. Specialists may, therefore, be considered *brittle* (or obversely, *robust*), if their performance greatly depends (or does not) on the competitive and trader context.

In this paper, we explore via simulation the extent to which the performance of specialists in the 2008 and 2009 Tournaments may generalise across competi-

tive and trader contexts. One of the aims of our analysis is to ascertain how it might be possible to create more robust specialists. Further, we wish to look at differences in overall performance of specialists submitted to the 2008 and 2009 competitions, to see what kind of progress is being made.

3 An Empirical Evaluation of the Generalisation Ability of the Entries

3.1 General Experimental Setup

Each JCAT simulation consists of a single tournament that runs for a number of trading days, with, in our experiments, 10 rounds per day and 500ms per round. All experiments were carried out with both the JCAT server and all clients situated on the same local machine. The trading population size was set at 400 traders, filled with traders taken from the four types described in Section 2. Buyers and sellers were split as evenly as possible in the different trader sub-populations. In order to achieve statistically significant results for each tournament variation, each was repeated 15 times, using the same configuration. The specialist agents we used in our experiments were downloaded in a pre-compiled form from the TAC Agent Repository². Specifically, we only used entries from the 2008 and 2009 competitions, except when this was not possible³.

As defined in Section 2, each day every specialist receives a score based on three criteria. A specialist's tournament score is the sum of its daily scores for all scoring days. Specialists are ranked in descending order of their total scores, with the specialist in rank 1 declared the winner of that tournament. To show that some specialists' performances can be sensitive to a number of factors, and in some cases generalise poorly, we measure differences in specialist performance across different tournament variations. Specifically, we look for two differences: firstly we measure the *qualitative impact* tournament variations have on each specialist's performance, which we achieve by comparing rankings of specialists' mean scores for different (comparable) tournament configurations. Secondly, we measure the *performance impact*, i.e., the change that tournament variations have on each specialist's score. It is clear that since the differences introduced between tournament variations, e.g., the proportions of trader types in the trading population, itself contributes to the performance of the specialists, one cannot simply compare the mean scores of specialists over tournament variations and confidently state how the specialists are able to generalise between the two cases.

Our response to this challenge is to define a new statistic to measure the performance of one specialist relative to others, across a diversity of tournament variations. Our statistic, which we call the normalised performance delta of a specialist, denoted $\hat{\delta}$, provides a metric for analysing how a given tournament

² <http://www.sics.se/tac/showagents.php>

³ Binaries not used were either unavailable, or had issues affecting their ability to be used in experiments, such as execution or library problems.

configuration affects the performance of the specialist. To calculate this statistic, we first calculate for each specialist i , the normalised mean score $\hat{\mu}_i$:

$$\hat{\mu}_i = \frac{\mu_i}{\sum_{j=1}^m \mu_j}$$

For a single specialist i , given two normalised scores $\hat{\mu}_i^x$ and $\hat{\mu}_i^y$ from two tournament variations x and y , we can calculate the absolute difference d_i^{xy} between the two scores:

$$d_i^{xy} = |\hat{\mu}_i^x - \hat{\mu}_i^y|$$

d_i^{xy} gives us an understanding of how, with respect to other specialists in the tournament, a specialist i 's performance has changed from one tournament variation to the next. Finally, for each specialist i we calculate the normalised difference value $\hat{\delta}_i$.

$$\hat{\delta}_i = \frac{d_i^{xy}}{\sum_{j=1}^m d_j^{xy}} \quad (1)$$

In order to ascertain some statistical significance to specialist mean score values generated from multiple tournament runs, two-tailed paired t-tests of equality of means were performed on certain pairs of specialists, in order to attempt to identify whether the reported rankings were distinct. In such cases both the *t-value* and *p-value* (using $n - 1$ df.) are reported. These tests assume normality in the distribution of the underlying random variables. However, even if the variables are from other distributions, the tests are still approximately correct [11] (p.197–198); [12] (p.362).

3.2 Evaluation of 2008 competition

For the analysis of the 2008 competition, the following specialists were included in our experiments: `CrocodileAgent`, `DOG`, `iAmWildCat 2008`, `Mertacor1`, `Mertacor2`, `PSUCAT`, `PersianCAT` and `jackaroo`. In the rest of this section we will often refer to them as `CR-08`, `DO-08`, `IA-08`, `M1-08`, `M2-08`, `PS-08`, `PC-08` and `JA-08` respectively.

Over-fitting to trading population In the following set of results, we show that some specialists' performances are sensitive to different mixes of trader types in the trader population, and as such some specialists may be over-fitted to specific types or mixes of traders. For this set of experiments, all of the specialists available were used. We ran a number of tournament variations, each of which consisted of differing proportions of individual trader types. In these tournaments, all 500 trading days were counted as scoring days. Overall, we found that several of the specialists' final rankings were affected by variations, particularly `Jackaroo`, `Mertacor1` and `Mertacor2`.

Table 1 shows the results of two tournament variations, which we refer to as ‘just-GD’ and ‘no-GD’. The just-GD variation consisted of a trading population made up of entirely GD traders, with equal buyers and sellers. In the no-GD variation, the trading population was composed of equal (as possible) proportions of RE, ZIP, and ZIC traders.

Specialist	Just GD Traders		No GD Traders		Rank	δ
	μ	σ	μ	σ		
PC-08	232.2	5.86	271.1	5.51	1, 1	0.355
M1-08	230.9	6.55	193.9	3.17	2, 4	0.364
JA-08	218.9	6.06	213.5	2.65	3, 3	0.064
M2-08	207.4	5.09	215.7	6.05	4, 2	0.067
IA-08	165.8	1.45	165.1	5.15	5, 6	0.016
DO-08	164.4	2.24	173.7	2.48	6, 5	0.078
CR-08	24.9	17.76	19.1	8.28	7, 7	0.057
PS-08	16.2	0.66	16.3	0.43	8, 8	0.001

Table 1: Mean, standard deviation, rank and δ values for a set of tournaments with just GD traders and a set of tournaments with no GD traders.

For a typical tournament variation, we found that in each of the n repetitions, scores, and thus rankings, were quite similar, leading to low σ values. It is extremely unlikely that scores would ever be identical over all runs due to the stochastic nature of the JCAT environment. Table 1 highlights the fact that the overall rankings for the two tournament variations were different, most notably with changes in the middle and lower portions. Qualitatively, we found of particular interest was the change in rank between M1-08, M2-08 and JA-08. In the just-GD case, M1-08 was rank 2 and M2-08 rank 4, while in the no-GD case the ranks were swapped to 4 and 2 respectively. In the just-GD case, a paired t-test showed that the average scores of M1-08 and M2-08 were significantly different, with a t -value of 9.36 and a p -value < 0.0001 . The average scores of M1-08 and M2-08 were 230.9 and 207.4 respectively. In the no-GD case, the t-test resulted in a t -value of 14.04 and a p -value < 0.0001 . Average scores in the no-GD case were 193.9 for M1-08 and 215.7 for M2-08.

Further, in the just-GD case we found that M1-08 and JA-08 had ranks of 2 and 3 respectively, while in the no-GD case they had ranks of 4 and 3. In the just-GD case, for M1-08 and JA-08, a t-test resulted in a t -value of 19.09 and a p -value < 0.0001 , with mean scores of 230.9 for M1-08 and 218.9 for JA-08. In the no-GD case, a t-test reported a t -value of 4.20 and a p -value < 0.0001 , with mean scores of 193.9 for M1-08 and 213.5 for JA-08.

Finally, we note that even a simple change of the trading population, i.e., just-GD, can make a previous winner, PC-08, lose its winning edge. Statistically, PC-08 is not the clear winner in the just-GD case. A t-test of equality of means between PC-08 and M1-08 in the just-GD case showed a t -value of 0.46 and a p -value of 0.65, with mean scores of 232.2 for PC-08 and 230.9 for M1-08. A counterpart to this situation is the no-GD case, where PC-08 clearly outperformed

M1-08. Here the t -value was 38.31 and p -value < 0.0001 , with mean scores of 271.1 for PC-08 and 193.9 for M1-08. This highlights a situation where either PC-08 or M1-08 are particularly sensitive to the proportions of GD traders in the population. The δ values for PC-08 (0.355) and M1-08 (0.364) were considerably larger than those of the other specialists, showing a disproportionate change in performance over the two cases for both specialists.

In Table 2 we show the results of two different tournament variations: ‘just-ZIC’ and ‘no-ZIC’. In the just-ZIC variation we used all specialists and a trader population containing only ZIC traders (equal number of buyers and sellers). The no-ZIC variation used all specialists, but the trader population contained an equal number—where possible—of GD, RE and ZIP traders.

Specialists	Just ZIC Traders		No ZIC Traders		Rank	δ
	μ	σ	μ	σ		
PC-08	276.2	11.47	244.1	4.53	1, 1	0.419
JA-08	218.6	4.71	214.6	2.49	2, 3	0.071
M2-08	213.9	8.20	224.7	5.77	3, 2	0.109
M1-08	192.1	2.72	207.0	3.76	4, 4	0.160
DO-08	170.7	2.62	180.0	2.32	5, 5	0.095
IA-08	161.6	5.15	170.7	4.09	6, 6	0.094
PS-08	16.9	0.46	16.2	0.47	7, 8	0.010
CR-08	16.8	5.07	20.4	7.90	8, 7	0.041

Table 2: Mean, standard deviation, rank and δ values for a set of tournaments with just ZIC traders and a set of tournaments with no ZIC traders.

We found another situation when it is not statistically clear that the rankings between two specialists are the same across the two tournament variations, indicating that there were generalisation problems. In the no-ZIC case, we found that M2-08 (rank 2) outperformed JA-08 (rank 3). A paired t-test of equality of means found the scores statistically different, with a t -value of 5.83 and a p -value < 0.0001 . The mean scores were 224.7 for M2 and 214.6 for JA. However, in the just-ZIC case we found that the mean scores between the two specialists, and thus the rankings, were not statistically distinct. A paired t-test results in a t -value of 1.75 and a p -value of 0.10. The mean scores were 213.9 for M2-08 and 218.6 for JA-08.

Table 2 also highlights a clear example of the performance impact that changes in the trader population can have on a specialist. The mean score for PC-08 in the justZIC case was 276.2, yet this dropped 8.83% to 244.1 in the noZIC case. Of course, the makeup of the trader population can have a significant impact on the scores of specialists, but in these two cases we see that other specialists’ scores did not vary proportionally as much as PC-08’s. By considering the performances changes of the other specialists, PC-08’s normalised delta value δ was 0.419, which was considerably higher than the others’.

We highlight in Table 3 other qualitative impacts that changes in the trader population had on specialists. In these tournament variations we considered the

cases ‘just-ZIP’ and ‘no-ZIP’. In the just-ZIP variation we used all specialists and a trading population consisting of only ZIP traders (equal number of buyers and sellers). In the no-ZIP variation all specialists were used but the trading population contained equal numbers—where possible—of GD, RE and ZIC traders.

Specialists	Just ZIP Traders		No ZIP Traders		Rank	δ
	μ	σ	μ	σ		
PC-08	255.7	12.18	255.9	5.17	1, 1	0.051
JA-08	231.3	5.02	211.5	4.37	2, 3	0.281
M1-08	208.0	5.86	199.0	4.19	3, 4	0.150
M2-08	189.0	5	221.9	6.94	4, 2	0.347
IA-08	169.7	3.16	173.5	4.34	5, 6	0.008
DO-08	164.1	4.86	179.7	2.4	6, 5	0.147
CR-08	17.0	6.35	16.1	5.04	7, 8	0.014
PS-08	16.2	0.65	16.3	0.4	8, 7	0.003

Table 3: Mean, standard deviation, rank and δ values for a set of tournaments with just ZIP traders and a set of tournaments with no ZIP traders.

In the just-ZIP case we found that M1-08 (rank 3) outperformed M2-08 (rank 4). A t-test resulted in a *t-value* of 10.35 and a *p-value* < 0.0001 . The mean scores for M1-08 and M2-08 were 208.0 and 189.0 respectively. However, in the no-ZIP case we again see a different outcome, with the ranks changed to 2 for M2-08 and 4 for M1-08. In this case, a t-test resulted in a *t-value* of 9.56 and a *p-value* < 0.0001 , with means of 221.9 for M2-08 and 199.0 for M1-08.

Over-fitting to other specialists In this set of results, we show that when the proportions of traders in the trader population remain fixed, some specialists’ performances are sensitive to the presence of other specialists in the marketplace. For this set of experiments, we used a fixed trader population containing an equal mix of GD and ZIC traders. Since ZIC traders are the least, and GD the most, sophisticated trader types, using this mix may offer the most diverse trading, and hopefully challenging environment. In these tournaments, all 500 trading days were counted as scoring days.

In Table 4 we see the effects that removing the bottom five specialists had on the remaining three. We particularly note the effect that lower-ranked specialists had on the performance of JA-08 and M2-08. For example, when all specialists were present, JA-08 outperformed M2-08 (*t-value* = 4.12, *p-value* = 0.0010). Mean scores for JA-08 and M2-08 were 217.5 and 206.9 respectively. Alternatively, when the lower five specialists were removed, and the remaining three compete with the same traders, we found that the rankings of JA-08 and M2-08 switched (*t-value* = 8.30, *p-value* < 0.0001). Mean scores for JA-08 and M2-08 were 243.5 and 256.72 respectively.

In Table 5 we can observe the effect that *removing* the top three specialists had on the remaining bottom five. In a qualitative context, IA-08 did significantly better than DO-08 when the top three specialists were not present, with a t-test revealing a *t-value* of 5.80 and a *p-value* < 0.0001 . The average score for

Specialist	All Specialists		Just PC, JA & M2		Rank	$\hat{\delta}$
	μ	σ	μ	σ		
PC-08	267.3	5.99	299.7	6.96	1, 1	0.361
JA-08	217.4	4.44	243.5	3.51	2, 3	0.293
M2-08	206.9	6.44	256.7	5.82	3, 2	0.347
M1-08	198.3	2.79	–	–	4, –	–
D0-08	174.4	2.81	–	–	5, –	–
IA-08	170.5	3.51	–	–	6, –	–
CR-08	18.8	5.21	–	–	7, –	–
PS-08	16.0	0.29	–	–	8, –	–

Table 4: Mean, standard deviation, rank and $\hat{\delta}$ values for a set of tournaments with all specialists and a set of tournaments with just PC, JA and M2. The trader population consisted of an equal mix of GD and ZIC traders.

IA-08 was 215.0, while D0-08 scored 208.7. However, when the top three specialists were present, D0-08 ranked higher than IA-08 (t -value = 3.31, p -value = 0.0051). The average score for IA-08 was 170.6, while D0-08 scored 174.5.

With respect to the performance impact that the two tournament variations had on specialists, it is clear that the inclusion (or likewise, exclusion) of the three specialists PC-08, JA-08 and M2-08 clearly affected the performance of M1-08 more than any of the remaining four. When the top three specialists were introduced, all of the other specialists' scores were lower, however a high normalised delta value of 0.525 for M1-08 indicated it was considerably more sensitive to their presence.

Specialist	All Specialists		No PC, JA or M2		Rank	$\hat{\delta}$
	μ	σ	μ	σ		
PC-08	267.3	5.99	–	–	1, –	–
JA-08	217.5	4.44	–	–	2, –	–
M2-08	206.9	6.44	–	–	3, –	–
M1-08	198.3	2.79	376.3	4.17	4, 1	0.525
D0-08	174.5	2.81	208.7	1.5	5, 3	0.198
IA-08	170.6	3.51	215.0	3.86	6, 2	0.217
CR-08	18.9	5.21	32.9	10.22	7, 7	0.044
PS-08	16.1	0.29	18.2	0.32	8, 8	0.016

Table 5: Mean, standard deviation, rank and $\hat{\delta}$ values for a set of tournaments with all specialists and a set of tournaments with all specialists except PC-08, JA-08 and M2-08. The trader population consisted of an equal mix of GD and ZIC traders.

Over-fitting to scoring period Our final results show that some specialists' performances were affected by the choice of trading days used as scoring days. In Sections 3.2 and 3.2 we maintained a fixed scoring period of 500 days and varied either the specialist or trader populations. In these results we take simulations from the previous sections, and adjust the scoring period from which the specialists' scores are generated. Specifically, we measured performance over the scoring day period 1–100 and 100–380.

Table 6 shows the effect changing the scoring period had on the performance of the specialists. In these two variations we used all specialists and a trading population containing just ZIC traders, with an equal number of buyers and sellers. For all specialists, as you would expect, scores for days 1–100 were much lower than the 100–380 tournament because the 1–100 scoring period is less than a third shorter. We found that many of the specialists’ ranks changed between the two cases, and highlight specifically JA-08 and M2-08. If specialists were ranked according to their score from the earlier period, we found that JA-08 had a rank of 3, while M2-08 had a rank of 2 (t -value = 10.42, p -value < 0.0001); mean scores were 46.0 for M2 and 39.8 for JA. Alternatively, if specialists were ranked using the scoring period 100–380 the ranks of M2 and JA switched. Again, scores were statistically distinct, resulting in a t -value of 3.34 and a p -value of 0.0048.

Specialist	Day 1–100		Day 100–380		Rank	δ
	μ	σ	μ	σ		
PC-08	50.8	1.94	158.6	7.34	1, 1	0.190
M2-08	46.0	1.86	118.8	5.19	2, 3	0.025
JA-08	39.8	0.81	124.4	3.34	3, 2	0.150
M1-08	39.5	1.01	107.7	1.80	4, 4	0.056
IA-08	35.0	1.19	89.1	3.72	5, 6	0.011
DO-08	34.0	1.09	96.4	1.46	6, 5	0.069
PS-08	16.9	0.46	0.00	0.00	7, 8	0.251
CR-08	16.8	5.07	0.00	0.00	8, 7	0.249

Table 6: Mean, standard deviation, rank and δ values for a set of tournaments with scoring days 1–100 and a set of tournaments with scoring days 100–380. The trader population consisted of just ZIC traders.

3.3 Evaluation of the 2009 competition

For the analysis of the 2009 competition, the following specialists were included in our experiments: Cestlavie (CE-09), Cuny.cs (CU-09), Jackaroo (JA-09), Mertacor (ME-09), PSUCAT (PS-09), TWBB (TW-09) and iAmWildCat 2009 (IA-09). As with the 2008 entries, we subjected the specialists to a variety of tournament configurations. Due to space considerations, it is only possible to present a small portion of these results in detail. Overall, we found that Jackaroo was a worthy champion, winning almost 80% of the tournament variants we tried, although there was a considerable lack of generalisation in the other specialists.

However, there were some configurations that JA-09 was particularly sensitive to. Along with CE-09, particular sensitivity was shown to GD traders. Table 7 clearly shows that JA-09 performs relatively poorly against just GD traders, while without their inclusion it regains its top spot. Paired t-tests of equality of means between JA-09 and CE-09 reveal t -value = 52.5 for just-GD, t -value = 24.47, and p -values < 0.0001 in both cases. CE-09 also dominated in the GD/ZIC mix variant, so it is perhaps the case that CE-09 has been engi-

Specialist	Just GD Traders		No GD Traders		Rank	δ°
	μ	σ	μ	σ		
CE-09	248.5	4.32	177.1	4.50	1,4	0.279
CU-09	200.5	6.15	158.6	6.86	2,6	0.171
JA-09	173.2	2.88	234.2	6.97	3,1	0.185
IA-09	168.5	2.84	160.8	3.22	4,5	0.049
ME-09	161.2	5.72	200.0	6.47	5,2	0.111
PS-09	138.6	9.15	178.4	2.52	6,3	0.117
TW-09	118.7	10.30	148.2	9.35	7,7	0.085

Table 7: Mean, standard deviation, rank and δ° values for a set of tournaments with just GD and no GD traders.

neered specifically to perform well against GD traders, thus generalising poorly against other types.

3.4 Evaluation of best performers and overall progress

In this section we take, based on previous experiments, the top three specialists from each year, and evaluate their performance. We considered previous specialist performances and took the three specialists with the highest mean scores. The specialists chosen from those entered in the 2008 competition were **Persian Cat**, **Mertacor** and **Jackaroo**. From the 2009 competition entries, **Cestlavie**, **Jackaroo (09)** and **Mertacor (09)** were chosen. Overall, we found that **JA-09** was

Specialist	Just RE Traders		No RE Traders		Rank	δ°
	μ	σ	μ	σ		
PC-08	230.1	5.75	153.1	5.72	1,6	0.481
JA-09	215.9	4.69	212.9	5.32	2,1	0.018
CE-09	189.4	3.10	189.5	4.43	3,4	0.001
ME-09	174.3	3.48	202.6	10.09	4,2	0.177
JA-08	169.0	3.25	178.4	3.91	5,5	0.059
ME-08	160.8	5.87	202.5	13.53	6,3	0.261

Table 8: Mean, standard deviation, rank and δ° values for a set of tournaments with just RE traders and no RE traders using the top three specialists from the 2008 and 2009 competitions.

still generally the strongest specialist, though the performance of all specialists was much closer. Of particular interest however, were the two variations involving RE traders; results can be viewed in table 8. In the just-RE case, we found that **PE-08** maintained—as it did against 2008-only specialists—its top position, but in the absence of RE traders its rank—which was still top against 2008-only competition—plummeted to 6th. In the just-RE (no-RE) case, a paired t-test of equality of means between **PC-08** and **JA-09** returned a *t-value* of 6.34 (47.33) and, in both cases, a *p-value* < 0.0001. By using our methodology to evaluate the generalisation properties of specialists we have discovered that a once winning specialists can over-fit to previously unseen competitors, which would have been

hard to identify without using such a methodology. This result is further interesting because in the 2009 only experiments, JA-09 won both the just-RE and no-RE cases, while in this case, an entry from the previous year outperformed it, suggesting previous competitor strategies had not been considered enough in the design of a new one.

One interesting question to ask is: what kind of overall progress is being made by specialists in general each year? In order to answer that we looked at the mean performance of specialists on different trader configurations for the 2008 and 2009 entries, as well as the 2008/9 best performers together.

Trader Mix	2008 Competition		2009 Competition		2008/9 Best	
	μ	σ	μ	σ	μ	σ
	Just GD	157.59	2.28	172.74	1.72	187.33
No GD	158.56	1.32	179.65	2.67	186.43	2.02
Just RE	160.52	2.25	182.00	1.22	189.96	1.98
No RE	158.13	0.66	181.03	2.09	189.87	1.66
Just ZIP	156.37	1.42	167.24	2.64	179.56	2.49
No ZIP	159.21	0.90	181.37	2.08	190.54	1.35
Just ZIC	158.35	0.97	179.72	1.80	188.16	2.11
No ZIC	159.71	1.14	181.18	2.13	189.68	1.59
GD/ZIC	158.75	0.89	180.33	1.79	191.24	1.59

Table 9: Mean specialist performance from tournaments with various trader configurations.

From Table 9 it is clear that specialist performances against any of the trader configurations were on average higher for the 2009 entries than the 2008 entries, and that when only the best performers from the two years were evaluated together, the mean performance was even higher. This suggests *some* progress is being made regarding the robustness of the designed specialists.

Perhaps the most interesting finding, which would not have been visible without such extensive experimentation, was that in all cases the trader population consisting of just ZIP traders invoked the lowest mean scores. Intuitively, one might expect that GD traders, using a more sophisticated strategy than ZIP, would be harder to perform well against. In all cases except for the 2008 experiments, a paired t-test returned p -values < 0.0001 (2009 t -value = 6.66, 2008/9 t -value = 6.53).

3.5 Discussion

We have presented in this paper a methodology that allows us to evaluate and compare agents' generalisation abilities systematically and quantitatively. We believe that this approach is the first to allow such properties to be measured—especially in a coevolutionary context. Further, there are several reasons why such an approach is important. Firstly, without such an approach it is hard to know the true performance of any particular agent strategy, as one cannot see how well a strategy performs against unknown or previously unseen competing

strategies, i.e., how well a given strategy generalises and thus its ‘robustness’. Although some of the specialists generalise better than we initially thought, our results show that some seemingly strong strategies can have weaknesses when competing with certain other strategies. This is interesting because it is clear from the literature that often ‘best strategies’ or ‘best results’ are published, and without any further elaboration on their generalisation ability, these results can be very misleading. We have also shown that a seemingly less intelligent trading strategy such as ZIP can be harder to perform well against than a more intelligent one such as RE or GD. Secondly, although it may seem obvious that the performance of any given strategy—whether in the CAT game or some other competitive multi-agent system—will depend on the strategies being used by other agents, it is *never* clear how strong this dependency is, or the resulting quantitative values such a dependency provides. Finally, we note that our approach’s significance can be demonstrated by emphasising that our results concur with the assumption that the performance of agents will depend on the strategies being used by their competitors, thus it is a sensible one.

Previously, Vetsikas and Selman [13] presented a methodology for deciding on the best strategy that their agent, **WhiteBear**, should use in the TAC classic game⁴. The authors decompose the overall bidding-agent problem into a smaller sub-problems, e.g., decide on the quantity of a good to buy, and then determine the best strategy given those quantities. One of the reasons that TAC is so hard for an agent is that there are multiple concurrent auctions taking place using a variety of rules. Their approach was to tackle each auction independently by generating a set of boundary strategies for each auction or good (an example of a boundary strategy might be ‘bid-low’ or ‘bid-high’). From these partial strategies they then generated intermediate strategy variants, which lay between the boundary strategies. Multiple experiments were carried out to test different strategies in the presence of differing numbers of other strategies. Although this approach is sensible when trying to decide on the best strategy from an initial set of possible strategies, it is unclear how well such a best strategy would fare against unseen strategies, i.e., other competitors in the trading competition. Indeed, unlike the approach in this paper, theirs provides no such mechanism to test the robustness of a strategy in the presence of unseen competitors.

Wellman et al. [14] explore the idea of using a reduced strategy space to allow them to evaluate potential strategies taken from their TAC agent, **Walverine**. They hypothesise that although a strategy’s performance depends on the strategies being used by other competitors, it may be relatively insensitive to the number of competitors using identical competing strategies. To that end [14] introduce a methodology that explores a reduced number of possible profiles—where a profile defines the type of strategies in use in a population and their frequency. They find that by reducing the strategy space, they can efficiently search the remaining space to find the strategies that can outperform the others. Although such an approach is very appropriate when considering a fixed initial space of potential strategies (in their case 35) as with the methodology in

⁴ <http://www.sics.se/tac/page.php?id=3>

[13] it is hard to say how well the found strategies would generalise to previously unseen ones, such as those faced in competitions.

Given the two previously discussed methodologies there are several reasons why our approach is novel. Firstly, the CAT game is more complicated than the original TAC game because there are two interacting populations—specialists and traders, both of whom contain (or could contain) members able to learn and/or evolve, and who are unknown in advance. Thus our approach is the first to look at a situation where there are *coevolving* populations of competing agents, and offer insights into the robustness of the strategies in those populations.

Secondly, to even define what ‘robustness’ of a trading or specialist strategy means is not straightforward in such a coevolutionary context, let alone to actually measure such a property in a systematic and rigorous way. Our approach, unlike the previously discussed ones, is the first to look at how well such designed strategies generalise to previously unseen strategies, using the CAT competition as a case-study. Indeed, we believe that our approach can be used for analysing the generalisation properties of strategies in not only other agent-based competitions, but in any agent-based simulation models within an complex adaptive domain.

4 Conclusions and Future Work

To our knowledge, the generalisation property of specialists (i.e., market mechanisms) has not been studied in the literature. Traditionally, economic mechanism design theory has dealt with the generalisation issue by seeking mechanisms which are incentive-compatible (i.e., which encourage truth-telling) and by assuming that all traders are always rational and self-interested. An incentive-compatible mechanism with rational, self-interested traders should generalise across trader strategies and types. However, because computer software is always resource-constrained and bug-prone, the designers of computational mechanisms cannot assume that traders always act rationally or in their own self-interest. Moreover, economic mechanism design theory has not considered competition between mechanisms, and so has not explored competitive performance of mechanism features, nor the generalisability of these features over different competitive environments. Within computer science, Niu *et al.* [15] considers the choices automated traders will make when faced with several competing online marketplaces, showing that although traders typically gravitate towards low-charging markets, such markets may lose their dominance when not all traders are experienced or do not all have access to full information.

Thus, the research reported here has tried to explore the generalisation properties of market mechanisms, using the 2008 and 2009 CAT Tournament specialists as the basis. It is unclear whether and how a market mechanism might (intentionally or unintentionally) favour certain trading strategies, or facilitate or inhibit other competing market mechanisms. A specialist which performs well under one particular tournament setup may not perform well if the setup changes slightly. Therefore, it is essential to study and understand any hidden bias that

a specialist might have built into it. In order to achieve this we ran many JCAT simulations using a variety of configurations, including changing both the trader and specialists populations as well as changing the period used to generate specialist scores. This paper shows for the first time how changes in the competition configuration can have an impact on specialists' performances in both a qualitative and quantitative context.

Our results showed that specialists can be sensitive (and specialised) to a number of factors in the competition, including trader strategies, other specialists, and the scoring period. Such results indicate that an appropriate evaluation of such a competition (and other similar ones) would need a theoretically sound framework, which can measure specialists' generalisation abilities quantitatively. They also point out the importance in analysing the relationship between a winning specialist and the particular competition setup used, so that insights into what makes a specialist better/worse can be gained. Further, we have shown that a trading strategy with less intelligence than others, i.e., ZIP, can be harder to score against, suggesting current specialists are not generalising well to it.

Although we studied three major issues that have a significant impact on specialists' performance from the viewpoint of generalisation, there are other issues to be considered, e.g., the performance metric used. It would be interesting to study the potential trade-offs between market share and profit, perhaps using a multi-objective approach. In terms of a theoretical framework for measuring specialists' generalisation quantitatively, we will investigate the possibility of adopting the one for measuring strategies' generalisation ability [16].

Finally, although the CAT Tournament aims specifically to encourage research and development of automated design of double auction mechanisms, we believe that the methods and techniques here have wider applicability. Indeed, many complex, adaptive domains, such as those in public policy or defence, have sub-populations of intelligent entities who co-evolve or change dynamically in response to each other's actions, quickly making any analytical mathematical models intractable. Accordingly, these domains are typically studied using computer simulation methods, and the issue of generalisability of results then becomes of great importance [17]. The techniques described in this paper potentially have application to these other domains, and we intend to explore such ideas in future work.

Acknowledgments

Supported by the Market Based Control of Complex Computational Systems project, EPSRC grants: GR/T10671/01 & GR/T10657/01. Financial support is gratefully received. We wish to acknowledge discussions and feedback from Peter Lewis, Michael Wooldridge, Tomasz Michalak, Andrew Dowell and Jinzhong Niu.

References

1. Gerding, E., McBurney, P., Niu, J., Parsons, S., Phelps, S.: Overview of CAT: A market design competition. Technical Report ULCS-07-006 Version 1.1, Depart-

- ment of Computer Science, University of Liverpool, Liverpool, UK (2007)
2. Gode, D.K., Sunder, S.: Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy* **101**(1) (1993) 119–137
 3. Cliff, D.: Minimal-intelligence agents for bargaining behaviors in market-based environments. Technical Report HPL-97-91, Hewlett-Packard Research Laboratories, Bristol, UK (1997)
 4. Erev, I., Roth, A.E.: Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88**(4) (1998) 848–881
 5. Gjerstad, S., Dickhaut, J.: Price formation in double auctions. *Games and Economic Behavior* **22**(1) (1998) 1–29
 6. Robbins, H.: Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58**(5) (1952) 527–535
 7. Niu, J., Cai, K., Gerding, E., McBurney, P., Parsons, S.: Characterizing effective auction mechanisms: Insights from the 2007 TAC market design competition. In Padgham, L.e., ed.: 7th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2008), NY, USA, ACM Press (2008)
 8. Petric, A., Podobnik, V., Grguric, A., Zemljic, M.: Designing an effective e-market: an overview of the CAT agent. In Ketter, W., ed.: Proceedings of 2008 Workshop on Trading Agent Design and Analysis (TADA 2008), Chicago, USA (2008)
 9. Vytelingum, P., Vetsikas, I.A., Shi, B., Jennings, N.R.: IAMwildCat: The winning strategy for the TAC market design competition. In Ghallab, M.e., ed.: 18th European Conference on AI (ECAI-2008), Patras, Greece, IOS Press (2008)
 10. Niu, J., Cai, K., McBurney, P., Parsons, S.: An analysis of entries in the first TAC market design competition. In Jain, L.e., ed.: IEEE-WIC-ACM International Conference on Intelligent Agent Technology (IAT 2008), Sydney, Australia (2008)
 11. Moses, L.E.: *Think and Explain with Statistics*. Addison-Wesley Publ. Co., Reading, MA (1986)
 12. Devroye, L.: *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY 10010 (1986)
 13. Vetsikas, I.A., Selman, B.: A principled study of the design tradeoffs for autonomous trading agents. In: AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems, New York, NY, USA, ACM (2003) 473–480
 14. Wellman, M., Reeves, D., Lochner, K., Cheng, S., Suri, R.: Approximate strategic reasoning through hierarchical reduction of large symmetric games. In: Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania, USA, AAAI (2005)
 15. Niu, J., Cai, K., Parsons, S., Sklar, E.: Some preliminary results on competition between markets for automated traders. In Collins, J., ed.: Proceedings of the AAAI 2007 Workshop on Trading Agent Design and Analysis (TADA 2007), Vancouver, Canada (2007)
 16. Chong, S.Y., Tino, P., Yao, X.: Measuring generalization performance in co-evolutionary learning. *IEEE Transactions on Evolutionary Computation* **12**(4) (August 2008) 479–505
 17. Marks, R.E.: Validating simulation models: a general framework and four applied examples. *Journal of Computational Economics* **30**(3) (2007) 265–290