TSA Administration

December 2003

A Preliminary Investigation

of

Question Bias

Initial Draft – September 2004

Contents

NOTE

This report is a draft that contains initial findings from the first bias analyses of TSA data. As discussed in the final section, there is much work still to be done before anything approaching final conclusions can be drawn.

TSA Administration December 2003

An Investigation of Question Bias

1 Introduction

The TSA administration in 2003 consisted of five tests given in the UK (Tests E, F, G, H and J) with two further tests that given overseas (Tests B1 and K). As part of the development of the UK testing, all tests were made available on-line but provision was made for Tests E and F also to be made available using paper and pencil. During this administration of TSA, the following numbers of applicants were tested.

	On-line	Paper	Total
Test E	81	441	522
Test F	98	485	583
Test G	89	0	89
Test H	86	0	86
Test J	83	0	83
All UK	**437**	**926**	**1363**
Test B1	0	28	28
Test K	0	160	160
All Overseas	**0**	**188**	**188**
All Applicants	**437**	**1114**	**1551**

In this, the third round of using TSA, the aim was to include an investigation of question bias in the analyses of the data collected. Specifically, a query hung in the air about the possible bias when using both paper and pencil delivery and on-line/on-screen delivery of the same test.

This report covers the main work on question bias that has been done and reports the main findings. In this work, the data used is that from the UK testing only (the 1363 applicants tested in Cambridge in December 2003) and does not include the data from the 188 students who were part of the overseas testing in China and Malaysia in October and November 2003.

Details of the scores achieved by applicants in various categories of the variables discussed herein may be found in the report by Raikes (2003).

2 Question Bias

Question bias is a term that is often used to mean different things by different people and is often used sloppily. So it may be helpful to start with a discussion of bias in order to understand the concept of bias before any results are presented.

> In mathematical statistics, "bias" refers to *systematic* under- or over-estimation of a population parameter by a statistic based on samples drawn from the population. In psychometrics, "bias" refers to systematic errors in the *predictive validity* or the *construct validity* of test scores of individuals that are associated with the individual's group membership. … It can involve any type of group membership – race, social class, nationality, sex, religion, age. The assessment of bias is a purely objective, empirical, statistical and quantitative matter entirely independent of subjective value judgements and ethical issues concerning fairness or unfairness of tests and the uses to which they are put. *Psychometric bias is a set of statistical attributes conjointly of a given test and two or more specified populations.* (Jensen, 1980, p. 375).

So, if we have two sets of , say Group A and Group B, who are differentiated by being members of two different population categories that are present and identified in the data from a TSA administration, what can we say if the test scores for Group A are significantly higher than those for Group B? At the heart of this question lies the real query: does the fact that in Group A scored much better than Group B, mean:

- that some of the questions in the test(s) used are biased towards the category defined by membership of Group A; or
- that those in Group A are simply better than those in Group B?

Clearly the answer to this question may well not be simple or clear cut since each factor may be present, to a degree. In looking at questions in an attempt to identify bias then, the aim is to identify systematic differences in group response patterns that occur, independent of the general level of performance.

To analyse the TSA data, an Item Response Theory (IRT) model, the one-parameter IRT model, the Rasch Model, was used (Rasch, 1960, 1980). To carry out the analyses, the software developed by the RUMM (Rasch Unidimensional Measurement Models) Laboratory, RUMM2020 (RUMM 2003), was used.

When using the Rasch Model the aim is to find a set of question and person statistics that explain as well as possible the data being analysed. The analysis thus derives optimal values for question and person statistics from the data using 'best fit' as a criterion. It may be, however, that for some questions the derived statistics (question difficulty and person ability) do not fully explain the response probabilities found in the data. If this is the case, then the question may be exhibiting DIF (Differential Item Functioning) where students from a specific category, e.g. Group A, generally find the question easier (or harder) than students in other categories, e.g. Group B.

The important point here is the concept of 'generally' as it is necessary to identify unusually easy or hard questions for one group of students relative to another and not to identify better or worse performance overall. There may well be genuine differences in performance between students in Groups A and B but the DIF analysis aims to identify questions that appear to be too easy or too hard for any particular group of students having controlled for any differences in overall level of performance of the groups.

Thus an investigation of bias is seen as a statistical analysis and the results of some of these analyses are what will be reported here. In the end, though, the purpose of doing the analyses is to be able to feed back to Question Writers what they should or, more likely should not do, to achieve questions that are bias free. As will be seen once the results are available, the various sets of questions may not give up their secrets easily. The interpretation of bias analyses in quite another matter from conducting them.

3 Looking at Differential Item Functioning (DIF)

To look at DIF it is necessary to collect data that allow the appropriate analyses to be conducted. In the case of the TSA in December 2003, six different aspects of bias were investigated. These six aspects are shown below.

3.1 Method of Delivery (Online)

A major question in the development and use of the TSA has been whether a parallel presentation of a test on paper and on screen would result in introducing bias of any kind. There were just two categories of delivery:

Delivery On-line		437
Delivery using Pencil and Paper	926	
All applicants	1363	

3.2 Gender

A common question for any major testing exercise is whether there is any gender bias. As the gender of students was collected as part of the registration this was included in the analyses. Information on gender was not available for one candidate.

Female	361
Male	1001
No gender given	1
All applicants	1363

3.3 Subject

All students were known to have a subject for which they were applying for admission to University. This variable was used to look at any potential bias that might exist across subjects. The subjects were as given below.

Computer Science	226
Economics	176
Land Economy	2
Mathematics	4
Natural Sciences	422
Engineering	533
All applicants	1363

3.4 Location

While the majority of applicants tested in the UK were from 'home' centres, some students came from an 'overseas' background. Testing students from a wide range background might well introduce issues of question bias and so this was investigated directly, the numbers of applicants being:

Overseas	214
Home	1138
Not Given	11
All applicants	1363

3.5 School Type

The possible effect of question bias arising from school type was another aspect of TSA that was seen as being very important to study. In the event, the following mix of school types was found in the sample of applicants tested:

Comprehensive		339
Sixth Form College	157	
Independent	452	
Grant Maintained	121	
Grammar	149	
FE	54	
Other	5	
Not Known	75	
Missing information	11	
All applicants	1363	

3.6 Decision

Finally, it was thought to be of interest to look at the two categories of applicants after a decision had been made to make a conditional offer or not. It is not an obvious source of bias but it was considered worthwhile to look at this aspect of the data.

Conditional Offer	411
Refer	918

No/missing decision	34
All applicants	1363

These six variables provide the basis for the bias analyses that follow.

4 Summary Results

The data were analysed using RUMM2020 and the DIF investigated for the six variables described above. The significance of the DIF for each question was looked at for the groups and any questions where there appeared to be significant differences between groups (i.e. where DIF appeared to be present) were noted.

The detailed results for the questions found to exhibit differential performance are presented in Appendix A. There the detailed results from the Analysis of Variance (ANOVA) that was conducted for each question can be found.

The ANOVA was carried out using students grouped into a maximum of 10 ability groups (referred to as class intervals in the analyses) and using the DIF categories. In each case, the significance of the main effect studied (DIF) is shown together with the significance of the particular variable being used and the interaction between the class interval and the variable. In common with usual ANOVA techniques, for a question to be judged as having significant DIF, the main effect needs to be significant, the variable needs to be significant and the interaction between the class interval and the variable needs not to be significant. If the latter is significant, then there is little basis for judging the significance of the DIF. An example of this argument is given below for the question that shows significant DIF in respect of the Method of Delivery variable. An example of a DIF plot is Given in Appendix B.

In a few cases, there was scope for some flexibility in interpretation of the ANOVA results for questions and this allows the identification of 'noisy' questions which, while not showing significant DIF, nevertheless show some consistent behaviour.

The following table provides a summary of the number of questions where significant DIF was found by category of question (CT or PS).

	Critical Thinking		Problem Solving		All Questions	
	Sig.	Noisy	Sig.	Noisy	Sig.	Noisy
Delivery mode	0	0	1	0	1	0
Gender	0	0	5	1	5	1
Subject	1	1	4	0	5	1
Location	6	2	6	0	12	2
School Type	1	1	0	0	1	1
Decision	0	1	1	0	1	1
All Variables	8	5	17	1	25	6

The five TSA tests used in the UK each consisted of 50 questions, potentially giving a total of 250 questions for analysis. As there were common questions between tests, however, there were only 180 unique questions used.
There were 25 questions that appear to be biased towards one group or another with a further six that, while not showing significant DIF were nevertheless decidedly 'noisy'. In fact there are three questions that each appear twice in the table so in fact only 22 questions are showing evidence of differential characteristics between the DIF categories.

Of these three questions, one will be seen to be noisy in two categories, one noisy in one category and significant in another and the third significant in two categories. No question appeared in more than two categories.

The next sections of this report discuss each of the DIF variables in turn and provide more detailed information on the nature of the biases.

5 The DIF Analyses

The analyses presented below provide details of the way in which questions were biased towards students in different categories within DIF variable. In each case the question label is given for reference and this is followed by an indication of the significance of the DIF found. The first figure relates to the significance of the DIF variable for the question and the second to the overall DIF for the question. Where there are many questions where a significant DIF was found, questions are presented in order of the overall DIF significance.

Appendix A provides the detailed analyses for each question discussed for each DIF variable.

5.1 Method of Delivery (Online)

The table below gives for each question where a query has been raised, details of the significance of the DIF found, the direction of the bias and a general description of the question concerned.

Question Label	Sig.	Bias To	Question Content
Oc0996016	0.00/0.00	P	Layout of tables/patterns

Note: P – bias towards paper and pencil version of question

The only significantly biased question in respect of Method of Delivery was highly significant and biased in favour of students answering in terms of paper and pencil rather than on-line. As may be seen, the main effect and the total DIF were highly significant. It can be seen from Appendix A that the interaction between the mode of delivery variable and the class intervals was not significant (p=0.11) so it is safe to infer that a significant DIF effect is present for the question as far as the mode of delivery variable is concerned.

The question provided students with pictures of a number of tables and then asked them to consider a number of possible re-arrangements. When presented with such a question on paper it would seem to be natural to sketch out possibilities and check thoughts on re-arrangement. With screen presentation, it would be more problematic to keep looking at the screen and then down at a piece of paper. It may also be the case that not all of the images in the question (original images of the tables and images in the options with different possible layouts) were able to be fitted on the screen at the same time thus requiring the students to scroll back and forth.

Conclusion It is probably wise to be careful administering questions that need visualisation of presented material using pencil and paper tests and on-screen assessment in the same session.

5.2 Gender

The table below gives for each question the information about the bias detected.

Question Label	Sig.	Bias To	Question Content
Oc0695010	0.00/0.00	M	Hockey league table – league results
Mo0991009	0.00/0.01	M	Clock faces – adjustment required?
Ms0393004	0.01/0.03	F	'Modern Art' model – painting
Ol0496007	0.03/0.04	M	Insurance claims table – best choice
Mo0195008	0.03/0.05	F	Combinatorial coffee machine problem
Oe0194003*	0.02/0.07	M	European birth-rate chart – selection

Note: F – bias towards females
M – bias towards males
* - a 'noisy' question

As indicated, five questions, as well as one noisy question, showed a gender bias. In three (plus one) cases, males achieved better than expected and females did better on two questions. In the case of the hockey league, it is tempting to argue that males are more commonly involved in football results where the scoring is similar, although not the same, to that in hockey. In the case of clock faces there is a strong visual element which is traditionally a weak area for females. The third question requires a logical thought process to work out the order of painting a model and there is no obvious reason why one gender should do better than the other.

The insurance claims question requires a degree of slogging through the presented facts and the fifth question requires a process of logical thinking to arrive at a position that allows the deduction of the answer. The final question, where the overall DIF was less significant than in the other cases, required reading from a chart to estimate a required statistic.

Conclusion There is little to go on here to provide any coherent interpretation of gender bias. It is possible that 'football'-type questions and questions requiring visualisation should be avoided and also that there might be a possible bias towards females in questions where logical thinking is required. There is, though, hardly sufficient information here to deduce that in these circumstances.

5.3 Subject

The table below gives for each question the information about the bias detected.

Question Label	Sig.	Bias To	Question Content
Dt0191001	0.00/0.00	N>C	'Most weaken' - advertising
Oc0196001	0.00/0.01	C>E	Mixtures – drawing/replacing
Of0694006	0.02/0.03	N<	Tennis knockout tournament
Mo0991008	0.04/0.04	X	Moon craters over time
Oc0695010	0.01/0.05	E<	Hockey league – league winner
Eo0794009*	0.00/0.07	C>E	'Main Conclusion' - vitamins

Note: N – Natural Sciences X – No clear pattern
C – Computer Science * - a 'noisy' question
E – Economics

The results here produce a very mixed set of observations and there is little that warrants building any firm conclusions. The first question, on mass advertising, was significant as the students applying fro Natural Sciences found the question relatively easier than those applying fro Computer studies. The second question, which required a logical approach to working out the proportions in a mixture, was found easier for the Computer Science Applicants than the Economics applicants.

A question on a tennis knockout championship was found relatively hard by Natural Science applicants but a question in the same general area, a hockey league, was found relatively difficult by those seeking to do Economics. This question is actually one found to be biased in favour of males (see above) who outnumbered females in Economics applications by 112 to 64. It cannot be clear, therefore, whether the bias for this question is a gender or a subject bias or, as is more likely, a mix of both.

The last (noisy) question, a logical deduction question, was found relatively easier by the Computer Science students than those applying for Economics.

Conclusion There is little consistent information to be drawn from these results.

5.4 Location

The table below gives for each question the information about the bias detected.

Question Label	Sig.	Bias To	Question Content
Eo0794009	0.00/0.00	H	'Main conclusion' – vitamins
Ds0790006	0.00/0.00	H	'Most weaken' – elephants and ivory
Ae0795195	0.00/0.00	H	'Underlying assumption' – lottery tickets
Ee1197091	0.01/0.02	O	'Parallel reasoning' – football matches
Mo0692013	0.03/0.02	O	International date formats
Md1094008	0.00/0.03	O	Folded card – spatial recognition
El0195030	0.02/0.03	H	'Strengthen' – teacher and eye contact
Mp1092002	0.00/0.03	H	Roundabout – numeric logic
Of1094008	0.01/0.04	O	Computational logic
Me1092005	0.00/0.05	O	Computational logic
Mb0291006	0.02/0.05	H	UK rail fares over time – inflation
Ee0497075	0.00/0.05	H	'Underlying principle' – business ethics
Cc0194068*	0.01/0.07	O	'Main conclusion' – early retirement
Ae0191109*	0.00/0.08	H	'Underlying assumption' – dental decay

Note: H – Home
O – Overseas
* - a 'noisy' question

The questions that were found to be significantly affected by DIF, and there were 12 (and two noisy questions as well), were split on whether they were found systematically easier/harder by Home or Overseas applicants. It must be remembered that the category of 'overseas candidate' does not refer to applicants from China and Malaysia who were tested overseas with TSA in Autumn 2003. The applicants in the Overseas group here were tested in the UK as part of the regular TSA testing. These are therefore students with an overseas background as opposed to a UK background.

As there is such a clear split in the questions, it is perhaps worth looking at the two groups of separately.

Questions Found Relatively Easier by Home than Overseas Applicants

The question on drawing a conclusion is short and to the point about vitamins and refers to eating fresh fruit and vegetables. This was also a 'noisy' question in respect of the subject variable so it is possible that there is more than one issue at stake here. The second question requires an ability to think widely from a given, short argument and the third requires rather more reading before an assumption might be drawn.

The question on the teacher in the classroom and eye contact may provide a scenario where those not familiar with UK teaching techniques feel unfamiliar; likewise the question on roundabouts, which is basically about logical thinking, may also provide an unusual scenario for some applicants.

The question on inflation of train prices is firmly UK based but the principle will not be unfamiliar to Overseas students. The last two questions are both quite wordy and the arguments presented convoluted but there is no apparent reason for any bias.

Questions Found Relatively Easier by Overseas than Home Applicants

The first question starts by using a football team analogy and requires clear thinking. The second is a question about international dates and it might seem reasonable that Overseas applicants should find such a question easier than Home applicants. The fact that the overseas category is related to applicants tested in the UK does, however, tend to weaken this point.

Then there is a three-dimensional problem with a net to be folded up, two questions requiring computational logic and one that requires the identification of the main conclusion of an argument.

Conclusion There is little consistent information to be drawn from these results although a question with an international perspective was found relatively easier by Overseas Applicants.

5.5 Decision

The table below gives for each question the information about the bias detected.

Question Label	Sig.	Bias To	Question Content
Mo0991001	0.00/0.01	C>R	Novel accounting system
Ae0191109*	0.00/0.06	C>R	'Underlying assumption' – dental decay

Note: C – received a Conditional Offer
R – referred
* - a 'noisy' question

Only one question showed up markedly here and was in favour of those students getting a conditional offer. The question required the understanding of a situation that was clearly stated but had an element of imprecision. It required an understanding of the use of approximation. The only other (noisy) question requires the understanding of an underlying conclusion to an argument. This latter question has already been seen as being 'noisy' in respect of the Location variable so once again a number of factors point to treating this question with caution.

Conclusion On a sample of one, and if the applicants who were made conditional offers are taken as the type of student who is wanted for admission to university, then perhaps questions with a 'novel approach' might elicit responses from those required for selection more readily than those who were referred.

5.6 School Type

The table below gives for each question the information about the bias detected.

Question Label	Sig.	Bias To	Question Content
En0293043	0.00/0.00	G>GM	'Most weaken' – Gambling in UK
Ah0191116*	0.04/0.07	I>	Re-cycling plastic bottles in US

Note: G Grammar
GM Grant Maintained
I Independent
* - a 'noisy' question

Again, only one question stood out when looking at this variable. A question requiring the understanding of an argument on gambling was found significantly easier by applicants from Grammar schools than for those from

Grant Maintained schools. A (noisy) question on seeking an underlying assumption from a discussion of re-cycling was found easier by applicants from Independent schools than by others.

Conclusion There is little consistent information to be drawn from these results.

In all that has been presented, there is little that really stands out as providing any evidence for consistent bias in TSA questions. While some questions have been identified as being significantly biased, much more work will be needed to try and identify real reasons for bias.

6 Summary and Further Work

Clearly the analysis of question bias reported here only relates to a single session of TSA and needs to be continued after forthcoming TSA test sessions if any coherent conclusions are to be drawn from the findings. In addition, the analyses might be extended in a number of ways.

For example:

The DIF analyses are dependent on the way that the data are divided up to form 'Class Intervals', or attainment groups. In the analyses reported her, 10 groups were used and it is clear that, for some questions at least, some of these groups were quite small. Repeating the analyses with fewer groups might serve to sharpen up the analyses and the subsequent interpretation of question bias.

From the current results, hypotheses can be formed that can be explored with the existing data. An example of this might be to classify questions according to criteria hypothesised from the results so far from 'biased' questions and then investigate the behaviour of 'similar' questions to see what signs of bias existed.

It would also be possible to include the applicants who were tested overseas and this might enable a different focus to be achieved on an interpretation of the 'Location' variable.

From all that has been found, however, it is clear that for the variables investigated there is no substantial and consistent bias in the case of most of the TSA questions used. In particular, apart from one question, no evidence of any difference between a paper and pencil administration and an on-line delivery has been found.

7 References

Jensen, Arthur, R. (1980). *Bias in Mental Testing.* London: Methuen.

Raikes, N. J. (2004). Preliminary Analysis of the Autumn 2003 TSA Tests. UCLES: ITAL Unit.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copanhagen: Danmarks Paedagogiske Institut.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.) Chicago: University of Chicago Press.

RUMM (2003). Rasch Unidimensional Measurement Models 2020 software. Duncraig, Western Australia): RUMM Laboratory Pty. Ltd.

Appendix A

ANOVA Results for Questions showing Bias

Item	Class Interval				**Online**				Online-x-CInt				Total DIF			
	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
--																
T0179	1.16	1.438	9	0.166575	11.63	14.435	1	0.000166	1.28	1.594	9	0.112048	23.18	2.878	10	0.001479

Item	Class Interval				**Gender**				Gender-x-CInt				Total DIF			
	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
T0880	1.00	1.096	9	0.362798	21.66	23.702	1	0.000000	0.67	0.732	9	0.679339	27.68	3.029	10	0.000870
T1379	1.08	2.588	9	0.012555	3.78	9.094	1	0.003605	0.84	2.027	7	0.064110	9.69	2.910	8	0.007494
T1048	1.94	2.711	9	0.004242	4.39	6.140	1	0.013508	1.12	1.561	9	0.123792	14.43	2.019	10	0.029513
T0855	1.17	1.335	9	0.215264	4.32	4.928	1	0.026829	1.40	1.593	9	0.113818	16.89	1.927	10	0.039278
T0802	0.85	0.856	9	0.568356	5.21	5.257	1	0.024951	1.64	1.659	7	0.134076	16.72	2.109	8	0.046659
T1084*	0.51	0.799	9	0.618479	3.91	6.176	1	0.015374	0.88	1.388	9	0.210607	11.81	1.867	10	0.065027

Item	Class Interval				**Subject**				Subject-x-CInt				Total DIF			
	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
T0502	1.17	1.224	9	0.276843	6.02	6.286	4	0.000049	1.28	1.331	28	0.120009	59.79	1.951	32	0.001540
T0831	1.31	1.505	9	0.142166	5.93	6.820	4	0.000021	0.88	1.012	28	0.449774	48.38	1.738	32	0.007755
T0646	0.59	0.632	9	0.770448	2.64	2.834	5	0.015417	1.23	1.322	29	0.122864	48.83	1.545	34	0.026933
T1378	0.63	0.863	9	0.563948	2.24	3.081	3	0.035898	1.17	1.609	21	0.086068	31.22	1.793	24	0.041737
T0880	1.00	1.088	9	0.368782	2.83	3.089	5	0.008970	1.08	1.172	31	0.238135	47.52	1.438	36	0.046579
T0107*	1.91	2.594	9	0.006168	2.55	3.462	5	0.004320	0.76	1.027	29	0.428682	34.71	1.385	34	0.074935

Item	Class Interval	**Location**	Location-x-CInt	Total DIF

	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
T0107	1.86	2.622	9	0.005627	22.80	32.204	1	0.000000	1.26	1.779	9	0.069195	34.14	4.822	10	0.000000
T0501	1.59	2.045	9	0.032277	15.59	20.074	1	0.000010	0.57	0.739	9	0.672837	20.76	2.673	10	0.003232
T0018	1.19	1.347	9	0.207937	14.68	16.674	1	0.000070	0.94	1.065	9	0.385889	23.12	2.626	10	0.003675
T0700	0.92	0.886	9	0.537144	6.24	6.026	1	0.014344	1.80	1.741	9	0.076569	22.45	2.169	10	0.018050
T1216	1.58	1.515	9	0.139357	5.15	4.932	1	0.026802	1.87	1.791	8	0.076405	20.12	2.140	9	0.024922
T0868	0.38	0.391	9	0.935806	9.65	9.820	1	0.002508	1.27	1.293	7	0.266278	18.54	2.359	8	0.026014
T0538	0.42	0.448	9	0.906529	5.01	5.314	1	0.022474	1.60	1.695	8	0.103525	17.81	2.097	9	0.032803
T1221	2.71	3.413	9	0.000446	9.41	11.837	1	0.000628	0.64	0.807	8	0.596623	14.54	2.033	9	0.034125
T0950	1.77	1.726	9	0.080400	6.66	6.485	1	0.011179	1.46	1.420	8	0.185071	18.32	1.983	9	0.039331
T1194	0.54	0.617	9	0.782760	7.00	8.071	1	0.004634	1.03	1.188	9	0.299757	16.28	1.876	10	0.045556
T1484	2.76	3.266	9	0.000660	4.50	5.319	1	0.021389	1.24	1.461	8	0.167923	14.39	1.890	9	0.050536
T0082	3.83	4.619	9	0.000027	9.15	11.018	1	0.000906	0.67	0.806	9	0.611199	15.16	1.827	10	0.051867
T0976*	1.66	1.607	9	0.109836	7.41	7.175	1	0.007602	1.12	1.080	8	0.375372	16.33	1.757	9	0.073596
T0003*	2.85	3.517	9	0.000286	8.71	10.739	1	0.001112	0.48	0.593	8	0.784083	12.56	1.720	9	0.081194

Item	Class Interval				**SchType**				SchType-x-CInt				Total DIF			
	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
T0788	6.08	5.350	9	0.000010	16.20	14.262	7	0.000005	0.85	0.748	50	0.898831	155.88	2.407	57	0.000000
T1254*	1.32	2.092	9	0.052808	1.46	2.300	7	0.044993	0.91	1.435	28	0.143401	35.63	1.608	35	0.072044

Item	Class Interval				**Decision**				Decision-x-CInt				Total DIF			
	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p	MS	F	DF	p
T1376	2.90	3.475	9	0.000352	9.22	11.057	1	0.000943	1.09	1.304	9	0.231483	19.00	2.280	10	0.012778
T0003*	2.91	3.615	9	0.000205	17.27	21.431	1	0.000016	-0.32	-0.394	9	**N/Sig	14.41	1.788	10	0.059580

Appendix B

An Example of Question Bias

The graph below shows an example of significant Differential Item Functioning for a TSA question. For the question (Oc0695010), it may be seen that the males did consistently better than the females throughout the range of ability.

This is a particularly clear case of bias but shows the nature of the differences that were found in the most significant questions.