



**An investigation of the use of the Angoff procedure
for boundary setting in multiple choice tests in
vocational qualifications.**

Report prepared by:
Sally Idle
Operational Research Team
OCR
July 2008

Abstract

The increasing use of e-assessment, in particular on-line multiple choice tests, in UK general qualifications (GCE and GCSE) has focused attention on the setting of cut scores, or grade boundaries, for the assessments. The Angoff procedure is a widely used procedure for estimating the difficulty of multiple choice items in Vocational Qualifications and thus setting cut scores for multiple choice tests.

Initially, subject matter experts (SMEs) are asked to make numerical estimates, in isolation, for the items which make up a test. They then discuss the items together and review the statistics on the actual performance of the items in the test. At this point they have the opportunity to revise their initial estimates. These revised item estimates are summed to give an estimated pass score for the test. The average of all the estimated pass scores from all the experts is then calculated to give the final pass score for the test.

By carrying out a meta analysis from the results of the Angoff procedure this paper investigates the accuracy of the SME's judgements and how they vary in the light of reviewing item statistics. By questioning whether such adjustments affect the validity of the Angoff procedure for multiple choice tests in vocational qualifications the generalisability of the findings to other types of qualification is considered.

Background

The Angoff procedure, developed by William Angoff in the 1970s, is an established and widely used procedure for setting cut scores for multiple choice tests. It has been used at OCR, in the context of vocational qualifications, for in excess of a decade.

The increasing use of on line multiple choice tests in UK general qualifications (GCE and GCSE) has focussed attention on the setting of cut scores or grade boundaries for such tests. A recent study by Novakovic (2007) used experimental data for an OCR vocational qualification and concluded that access to statistical data had a positive effect on the Angoff procedure. This investigation looks at actual data from the Angoff procedure as applied to two OCR vocational qualifications.

The OCR Angoff Procedure

Several weeks before the Angoff awarding meeting the awarders receive a pack containing question papers, answer keys, OCR procedures and instructions and a form to record their initial estimates. They are asked to estimate the percentage of minimally competent candidates (MCCs) that would get the question correct. (A minimally competent candidate is one who has covered all the syllabus for the test, has been well prepared and whose ability level is just sufficient to achieve a pass.) The awarders make their initial estimates and return them to OCR.

All the awarders then attend the Angoff awarding meeting which is chaired by an experienced OCR chairperson. They are given back their initial estimates and statistical data (item facility and discrimination values) on the actual performance of the items in the test. The chairperson then conducts the panel through the test, question by question, looking at the statistical data and offering opportunity for discussion. At the conclusion of the discussion for each question the panel members record their second estimates for the performance of MCCs, which may be different from the first.

After the meeting the second estimates are summed by awarder giving each awarder's estimate of the pass score. The mean of all the pass scores then goes forward as the estimated pass score for the test.

This Investigation

Data

The seven sets of data used in this investigation come from the results of the Angoff procedure as applied to multiple choice tests from two OCR vocational qualifications.

OCR Level 3 Certificate of Professional Competence (CPC) in National Road Haulage or National Passenger Transport is offered on behalf of the Department for Transport and is a requirement for the purpose of Operator Licensing. The mandatory core unit (unit 1) is a 30 question multiple choice test which is offered four times per year. The tests are constructed from 50% new and 50% previously used items. Four data sets are from sessions in 2006/07.

OCR Level 2 Award in Administration (AinA) is part of the Modern Apprenticeship programme in Administration. Unit 1 is a 30 question mandatory on-line multiple choice test of new items. Three data sets are from 2007.

In the awarding of these tests, one question was removed from each of CPC Test 1, AinA Test 1 and AinA Test3, so these had a maximum score of 29.

Awarders

The Angoff procedure has been used for both qualifications for four years so all the awarders are experienced and familiar with the procedure. The panel consists of subject matter experts who are drawn from people who set tests, people involved in teaching/training and, in the case of CPC, from the transport industry. It is not usual to have the setter of the test on its Angoff panel.

For the CPC tests there were a total of seven awarders, two tests had five awarders, one had six and one seven. Three were involved in all four tests.

For the Award in Administration tests there were a total of nine awarders, all three tests had five awarders. One of the awarders was involved in all three tests.

Analyses

Minimally competent candidates

It was necessary for comparison purposes to identify the MCCs who completed each test and calculate their facility values. The group of MCCs was defined as those candidates whose score was between plus or minus 1 SEM (standard error of measurement) of the pass mark for the test.

Table 1 records the total number of candidates for each test, the number identified as MCCs and the mean mark achieved by each.

Table 1 The average performance of all candidates and MCCs

	Number of Candidates	Number of MCCs	Mean mark (all candidates)	Mean mark (MCCs)
CPC Test 1	1759	638 (36%)	19.78 (66%)	19.05 (64%)
CPC Test 2	1775	636 (36%)	21.19 (71%)	20.08 (67%)
CPC Test 3	1652	579 (35%)	21.42 (71%)	20.05 (67%)
CPC Test 4	1332	452 (34%)	21.32 (71%)	20.19 (67%)
AinA Test 1	101	52 (51%)	20.81 (69%)	21.56 (72%)
AinA Test 2	131	74 (56%)	19.77 (66%)	19.15 (64%)
AinA Test 3	156	79 (51%)	22.39 (75%)	21.35 (71%)

In general the mean mark for MCCs is lower than that for the whole group. AinA Test 1 is the exception to this, the small number of candidates and the way the marks are distributed being responsible for this.

Agreement among awarders

The reliability between awarders was calculated by using the Intra Class Correlation Coefficient with a two-way random effects model (the awarder effects and the item effects are random). An ICC of 1 would indicate that there was perfect agreement between the awarders on each item. Table 2 shows the average ICC measures for all tests.

Table 2 Average ICC measures of all tests

	First estimates	Second estimates
	ICC	ICC
CPC Test 1	0.80	0.94
CPC Test 2	0.76	0.93
CPC Test 3	0.69	0.90
CPC Test 4	0.67	0.89
AinA Test 1	0.50	0.97
AinA Test 2	0.24	0.89
AinA Test 3	0.62	0.95

For the first estimates the inter-awarder reliability was good for all tests apart from AinA Test 2. In every test the reliability increased for the second estimates to very high values. This indicates a high level of agreement among the awarders.

Correlation

To measure the extent to which the awarders were able to rank the items in order of difficulty, both the first and second estimated facility values were compared to the actual facility values for minimally competent candidates using the Spearman rank-order correlation coefficient. These results are shown in Table 3. They were also compared to item difficulty measurements calculated from the actual statistics. These results are shown in Table 4.

Table 3 Spearman rank-order correlations between estimated facilities and actual facilities for MCCs

	First estimates	Second estimates
CPC Test 1 actual facilities (MCCs)	0.80	0.94
CPC Test 2 actual facilities (MCCs)	0.72	0.95
CPC Test 3 actual facilities (MCCs)	0.62	0.85
CPC Test 4 actual facilities (MCCs)	0.49	0.89
AinA Test 1 actual facilities (MCCs)	0.44	0.90
AinA Test 2 actual facilities (MCCs)	0.31	0.46
AinA Test 3 actual facilities (MCCs)	0.49	0.93

Table 4 Spearman rank-order correlations between estimated facilities and item difficulties

	First estimates	Second estimates
CPC Test 1 actual facilities (MCCs)	0.78	0.94
CPC Test 2 actual facilities (MCCs)	0.71	0.95
CPC Test 3 actual facilities (MCCs)	0.62	0.84
CPC Test 4 actual facilities (MCCs)	0.53	0.90
AinA Test 1 actual facilities (MCCs)	0.59	0.97
AinA Test 2 actual facilities (MCCs)	0.39	0.52
AinA Test 3 actual facilities (MCCs)	0.35	0.88

Both sets of correlation data show very similar results.

At first estimate stage, the correlation between estimated facilities and actual facilities for MCCs is better for CPC tests than for AinA tests. The correlation after second estimates is better in all cases and is very good for all tests except for AinA Test 2. In general all the awarders are good at ranking items in order of difficulty with the aid of discussion and actual statistics. CPC awarders are better than AinA awarders without these aids.

Accuracy of the awarders

Tables 5a and 5b show, for all tests, the frequency of changes for each awarder between the first and second estimates.

Table 5a Frequency of changes between first and second estimates (CPC)

	CPC Test 1	CPC Test 2	CPC Test 3	CPC Test 4
Awarder 1	22	21	19	18
Awarder 2	13	7	8	11
Awarder 3	10	9	4	4
Awarder 4	15	22		18
Awarder 5	15	17		18
Awarder 6		24	13	17
Awarder 7		22	17	
All Awarders	15.0	17.4	12.2	14.3

Table 5b Frequency of changes between first and second estimates (AinA)

	AinA Test 1	AinA Test 2	AinA Test 3
Awarder A	11	11	13
Awarder B	20		21
Awarder C	21		22
Awarder D	29		
Awarder E	21		
Awarder F		23	
Awarder G		13	
Awarder H		24	23
Awarder I		23	22
All Awarders	20.4	18.8	20.2

The tables show that the awarders make many changes to their first estimates. The average number of changes per awarder is higher for Award in Administration (67%) than for CPC (50%). Where awarders are involved in more than one of the tests, it can be seen that the number of changes made by an awarder is consistent from test to test.

Table 6 shows the recommended pass mark after the first and second rounds of estimates. These have been calculated by averaging the sum of the awarders' estimates for each test.

Table 6 The awarding panel's recommended pass marks after first and second estimates

	Mean mark (all candidates)	Mean mark (MCCs)	Pass mark recommended by awarders (First estimates)	Pass mark recommended by awarders (Second estimates)
CPC Test 1	19.78 (66%)	19.05 (64%)	19 (66%)	19 (66%)
CPC Test 2	21.19 (71%)	20.08 (67%)	21 (70%)	20 (67%)
CPC Test 2	21.42 (71%)	20.05 (67%)	20 (67%)	20 (67%)
CPC Test 2	21.32 (71%)	20.19 (67%)	20 (67%)	20 (67%)
AinA Test 1	20.81 (69%)	20.56 (72%)	20 (69%)	21 (72%)
AinA Test 2	19.77 (66%)	19.15 (64%)	20 (67%)	19 (63%)
AinA Test 3	22.39 (75%)	21.35 (71%)	20 (69%)	21 (72%)

From the data in Table 6 it can be seen that for CPC, although the awarders made many changes to their estimates, in three out of the four tests the recommended pass mark did not change between the first and second estimates. However, for Award in Administration, in all cases the recommended pass mark changed by one mark.

Table 7 shows whether awarders' estimates were lower than, higher than or close to the actual facility values for MCCs. The three categories used were underestimates, accurate estimates and overestimates as used in Goodwin (1999), Impara & Plake (1997) and Novakovic (2007). Accurate estimates were defined as those falling within ten percentage points above or below the actual facility values for MCCs, overestimates were those above this boundary and underestimates below.

Table 7 Accuracy of Awarders' estimates compared with the actual performance of MCCs

		Under		Accurate		Over	
		N	%	N	%	N	%
CPC Test 1	First Estimates	35	24	68	47	42	29
	Second Estimates	29	20	84	58	32	22
CPC Test 2	First Estimates	76	36	64	29	73	35
	Second Estimates	64	29	92	44	57	27
CPC Test 3	First Estimates	48	32	55	37	47	31
	Second Estimates	33	22	78	52	39	26
CPC Test 4	First Estimates	52	29	81	45	47	26
	Second Estimates	37	21	108	60	35	19
AinA Test 1	First Estimates	66	38	65	37	43	25
	Second Estimates	41	23	107	61	26	15
AinA Test 2	First Estimates	60	40	34	23	56	37
	Second Estimates	50	33	58	39	42	28
AinA Test 3	First Estimates	69	48	45	31	31	21
	Second Estimates	51	35	76	52	18	12

This table shows that for the first estimates the CPC awarders achieved less than 50% of accurate estimates with a roughly even split of under and overestimates. The AinA awarders had a lower percentage of accurate estimates than the CPC awarders with more underestimates than overestimates. For second estimates the percentage of accurate estimates increased in all cases. For CPC there was still an even split between under and over estimates and AinA there was still more underestimates.

The AinA awarders have a greater tendency to underestimate the performance of MCCs. It would appear from these data that there is consistency of awarders' behaviour within a qualification but not across the two qualifications.

Actual and absolute differences

The frequency table above gives information on the direction of differences between estimated and actual facilities for MCCs. They do not however provide information on size or frequency of the actual changes made by awarders. Unless awarders changed their estimate by a large enough amount to result in it being reclassified, for example from an underestimate to an accurate estimate, it would not be recorded in the frequency table. To investigate the magnitude of change between the first and second estimates the mean actual and mean absolute differences between actual and estimated facilities for MCCs were calculated for each awarder.

The actual differences were calculated by subtracting the actual facilities for MCCs from the estimated values. Positive values of the mean show that an awarder has mostly overestimated, negative values show that the awarder has mostly underestimated, while values close to zero could indicate either a high degree of accuracy or erratic estimating (even quantities of over and underestimation.) Comparing the mean actual differences for first and second estimates shows if awarders made significant changes to direction.

The absolute differences were calculated by subtracting the actual facilities for MCCs from the estimated values and making all differences positive. Mean absolute differences show the size of error in the estimates.

Tables 8a and 8b show the actual mean differences for all awarders for all tests.

Table 8a Mean actual difference between estimated and actual facilities for MCCs (CPC)

	CPC Test 1		CPC Test 2		CPC Test 3		CPC Test 4	
	First Est	Second Est	First Est	Second Est	First Est	Second Est	First Est	Second Est
Awarder 1	-8.19	-8.03	-5.93	-8.26	-7.99	-6.82	-11.29	-8.63
Awarder 2	9.64	6.31	6.57	3.74	4.68	4.34	2.71	1.87
Awarder 3	2.14	1.31	5.24	4.74	1.18	1.01	5.37	4.54
Awarder 4	4.14	2.47	7.07	4.57			0.21	1.71
Awarder 5	-1.69	-1.69	-2.10	-2.10			-4.79	-3.46
Awarder 6			8.57	4.90	4.01	5.01	3.97	4.04
Awarder 7			-8.26	-1.10	-0.16	3.34		
Mean	1.21	0.07	1.60	0.93	0.34	1.38	-0.64	0.01

Table 8b Mean actual difference between estimated and actual facilities for MCCs (AinA)

	AinA Test 1		AinA Test 2		AinA Test 3	
	First Est	Second Est	First Est	Second Est	First Est	Second Est
Awarder A	2.10	-0.06			-10.10	-3.60
Awarder B	-3.40	0.27			-5.77	-2.44
Awarder C	-7.90	-3.73				
Awarder D	-1.40	-0.23				
Awarder E	-7.40	-7.06	3.12	-2.88		
Awarder F	4.60	-0.06	10.62	6.12	-5.44	-7.44
Awarder G			-18.22	-15.88		
Awarder H			-6.05	-2.38	-2.44	-3.10
Awarder I			9.12	8.45	-9.10	-4.44
Mean	-2.23	-1.81	-0.28	-1.32	-6.57	-4.20

Table 8a shows that for CPC some awarders overestimate and others underestimate with the two effects cancelling each other out. CPC awarders are consistent in their behaviour from test to test. Table 8b shows that for AinA more awarders overestimate than underestimate with Test 3 having no awarders overestimating. Awarders are not consistent from test to test.

For all tests apart from AinA Test 2 the overall size of the error decreased between first and second estimates.

Figures 1a and 1b show graphically the change in the size of mean actual difference between estimates.

The graphs show clearly that in most cases the awarders' mean actual differences move close to zero. Generally there are no significant changes to the direction of estimate. In fact there is only one case (Awarder E, AinA Test 2) where the sign of the mean actual difference

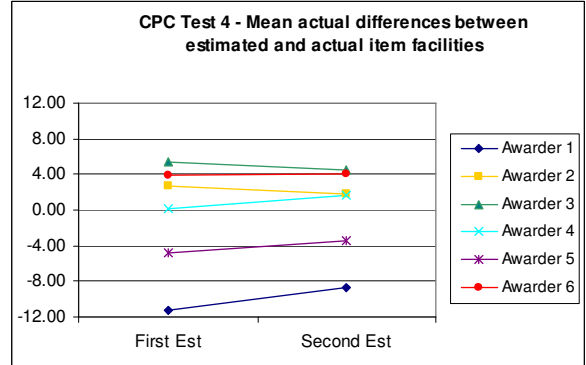
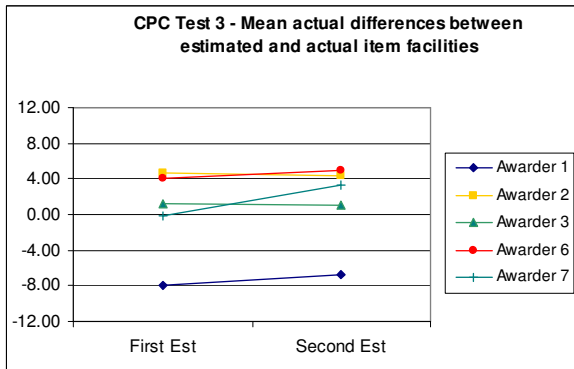
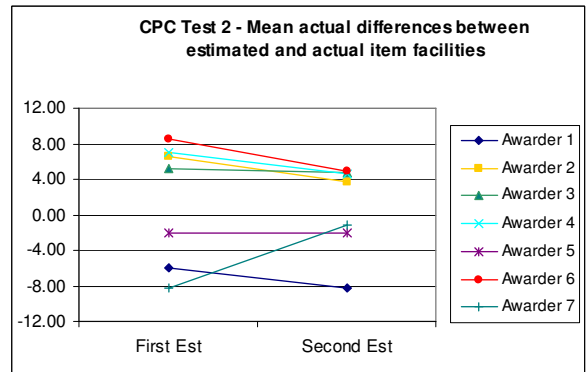
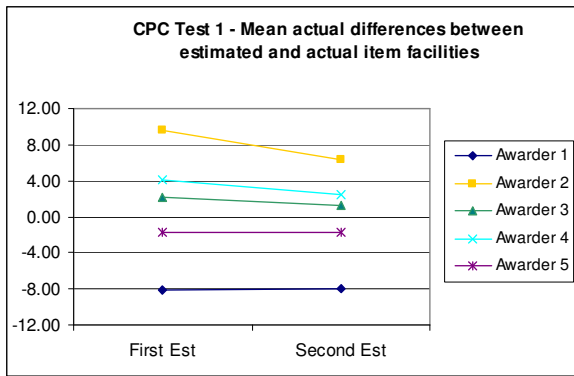


Figure 1a: The mean actual differences between estimated and actual facility values for MCCs for the CPC tests.

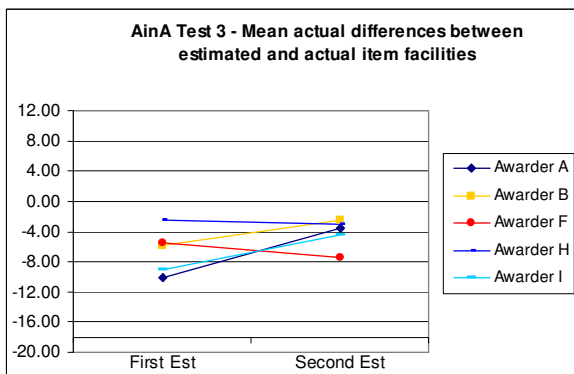
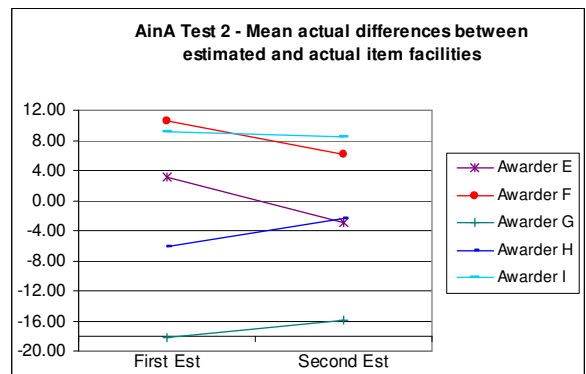
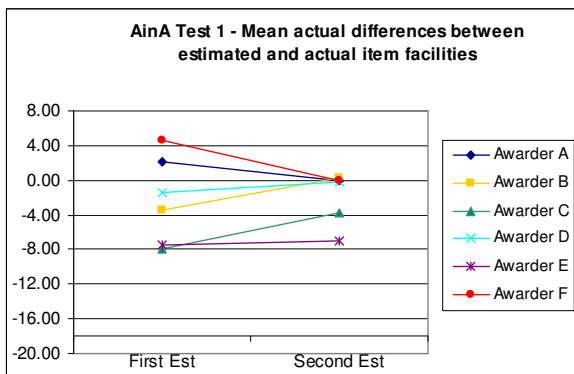


Figure 1b: The mean actual differences between estimated and actual facility values for MCCs for the AinA tests.

Tables 9a and 9b show the mean absolute differences for all awarders for all tests.

Table 9a Mean absolute difference between estimated and actual facilities for MCCs (CPC)

	CPC Test 1		CPC Test 2		CPC Test 3		CPC Test 4	
	First Est	Second Est	First Est	Second Est	First Est	Second Est	First Est	Second Est
Awarder 1	13.58	8.67	18.81	11.00	16.44	10.40	15.50	9.79
Awarder 2	13.56	9.66	15.37	14.73	15.10	13.10	11.08	8.58
Awarder 3	12.38	10.55	16.82	14.66	13.64	12.81	12.54	11.04
Awarder 4	12.26	8.54	15.74	8.85			13.77	7.94
Awarder 5	12.50	9.12	17.08	11.42			14.90	9.63
Awarder 6			17.67	14.05	15.55	11.87	15.86	12.44
Awarder 7			16.32	8.66	14.77	10.11		
Mean	12.87	9.31	16.83	11.91	15.10	11.66	13.94	9.90

Table 9b Mean absolute difference between estimated and actual facilities for MCCs (AinA)

	AinA Test 1		AinA Test 2		AinA Test 3	
	First Est	Second Est	First Est	Second Est	First Est	Second Est
Awarder A	16.10	9.73			18.62	9.40
Awarder B	16.96	10.55			16.87	8.18
Awarder C	20.18	9.09				
Awarder D	22.37	8.92				
Awarder E	15.63	11.35	19.45	12.76		
Awarder F	16.53	11.09	18.38	12.88	14.85	10.56
Awarder G			24.37	20.73		
Awarder H			22.38	13.08	17.36	10.69
Awarder I			22.04	14.71	19.31	12.69
Mean	17.96	10.12	21.33	14.83	17.40	10.31

The decrease in the mean absolute values between first and second estimates for all awarders in all tests indicates a decrease in the magnitude of error.

Generally the magnitude of error at the first estimate stage was higher for the AinA tests than for the CPC tests, but after the second estimates the level of error remaining was similar in all tests.

Figures 2a and 2b show graphically the change in the size of meal absolute difference between estimates.

Graphically it is clear to see the reduction in error for all cases.

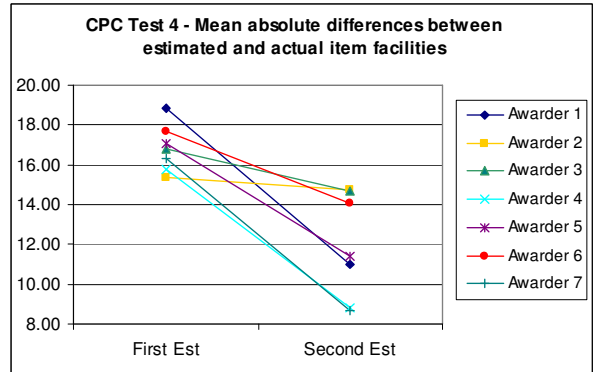
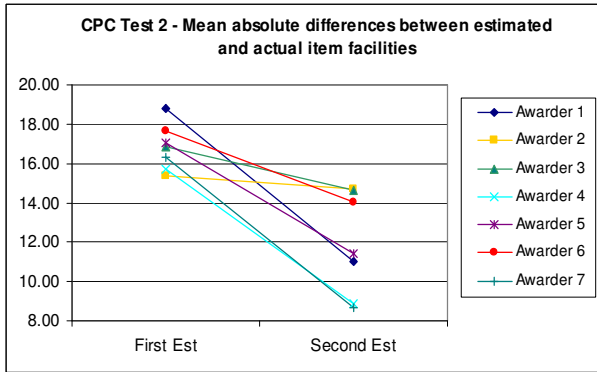
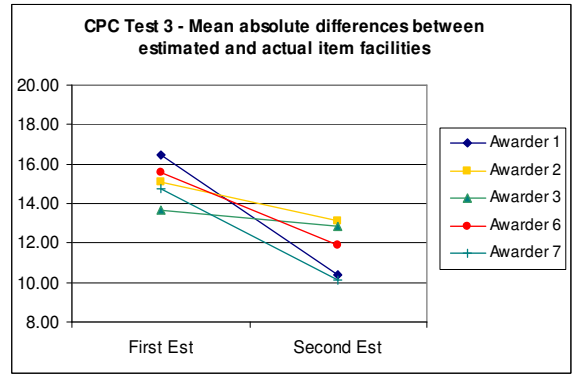
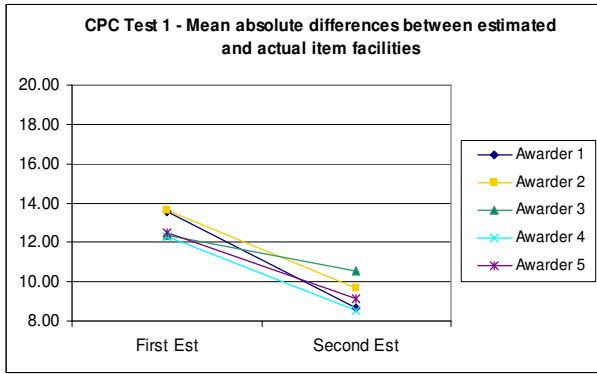


Figure 2a: The mean absolute differences between estimated and actual facility values for MCCs for the CPC tests.

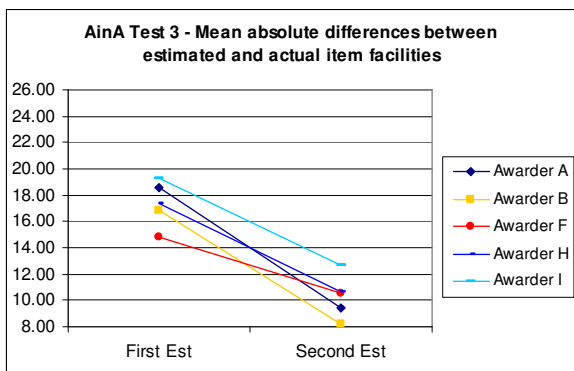
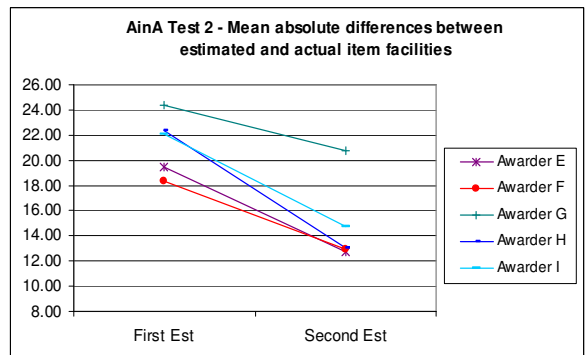
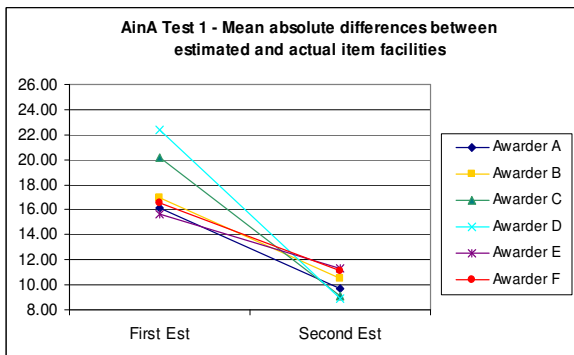


Figure 2b: The mean absolute differences between estimated and actual facility values for MCCs for the AinA tests.

Discussion

The results of this study have shown that the performance of awarders improves between making their first and second estimates. The increase in ICC shows an improvement in the inter-awarder reliability; the higher correlation coefficients show increased accuracy in ranking items in order of difficulty; the number of estimates classified as accurate when compared to the actual performance of MCCs increases; and the mean absolute differences show a decrease in the error of estimates when compared to the actual performance of MCCs.

The fact that the mean number of changes made is more than 40% for all tests shows that all the awarders were confident enough to make changes to their initial estimates. However the changes they made had very little effect on the recommended pass mark. This remained unchanged for four out of the seven tests and for two of the tests where it did change this was caused by a very small change (0.2 and 0.3 marks) which affected the rounding to a whole number.

It has been established that the second estimates are more accurate than the first, a fact which could be expected to lead to a more accurate recommended pass mark. That it has had very little effect on the pass mark can be put down to the effect of averaging estimates for the set of awarders. This is further reinforced by more detailed investigation of the correlation coefficients. When the individual correlation coefficients for each awarder were examined they were found to be very variable and in every case lower than the correlations for the averaged estimates.

The percentage of estimates classified as accurate in the first rounds of estimates was low (less than 40% in most cases). The level of accuracy increased after the second estimates, but still only reached 60 % for two tests and 50% or less for the rest. This means that for the majority of the tests studied as many of the estimates were inaccurate as were accurate. This calls into question the validity of the Angoff procedure in these tests, where the pass mark has been recommended on the basis of only half of the estimates being classified as accurate.

From this investigation differences between the two sets of awarders have been identified. The CPC awarders were generally more accurate than the AinA awarders at the first estimate stage and made fewer changes to their estimates. The AinA awarders had a greater tendency to underestimate the performance of MCCs in both first and second estimates.

This study has looked at the results of the Angoff procedure for seven live tests across two OCR vocational qualifications. It has raised questions about the percentage of accurate estimates and hence the validity of the Angoff procedure as applied to these qualifications. Further study, involving more tests and other qualifications would help to confirm the generalisability or otherwise of these results.

References

Goodwin, L. (1999) Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied measurement in Education*, 12:13-28.

Impara, J. & Plake, B. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34(4):353-366.

Novakovic, N. (2007). The Influence of Discussion and Statistical Data on Judges' Estimates during Angoff Awarding Meetings. *Cambridge Assessment*.