Exploration of Assessment Quality Issues in A Level Modern Foreign Languages

Technical Report

September 2014

Ofqual/14/5519

Contents

1 Introduction

In light of concerns around the appropriateness of standards in A level modern foreign languages (MFL), this paper explores potential issues with the quality and functioning of the assessments making up the current specifications. The specifications considered are the A levels in French, German and Spanish offered by AQA, OCR, Pearson and WJEC.

Examination of assessment quality is important to seek assurances that candidates are receiving marks, and ultimately grades, that represent their true ability. Further, inadequate assessment quality may underlie real or perceived grading standard issues. The objective of this work is to identify, where appropriate, both assessment design issues that should be addressed for the current specifications, and factors that should influence the design of future specifications and assessments as part of the upcoming reforms.

Two strands of activity have informed this report. First, a scrutiny exercise performed by subject experts to identify any potential problems with the quality of the assessments. Second, a technical exploration of the candidate-level data to identify any quantitative indicators that may or may not suggest issues with the technical functioning of the assessments. The common objective of these strands is to identify any aspects of the assessment designs that may compromise the validity of the assessment and, more specifically, the validity of the rank order of candidates. The outcomes of both strands of work are presented here alongside any recommendations for modification of assessment design or for further exploratory work. Individual candidate responses to the assessments were not considered as mark of this exercise.

Whenever one conducts analyses of the type reported here, it is highly likely that some issues will be identified. The production of assessments that function in a completely ideal manner is extremely challenging and is unrealistic in the absence of a robust and extensive programme of pretesting. This is reflected in the findings of this work.

2 Methodology

2.1 Subject expert review

Three subject experts per language were commissioned to carry out the review of assessment materials, with one of these experts acting as lead reviewer for the language. The assessment materials reviewed were the summer 2013 written papers, mark schemes and speaking tests for GCSE and A level qualifications. This report focuses on the A level findings.

Feedback from stakeholders, including those attending a teachers' conference hosted by Dulwich College in November 2013, suggested that there were a number of areas of concern with the assessment of languages. Before reviewing the materials, the expert reviewers attended a briefing meeting with a selection of teachers from each of the three languages, including some who had attended this conference. This was to give the reviewers a first-hand account of the specific concerns that language teachers had reported and to confirm the questions for investigation. The pre-briefing exercise gave the reviewers a particular focus. However, issues raised by the qualitative review were tested through quantitative data analysis so protecting against the introduction of any potential bias.

The questions for investigation were confirmed and focused on the following areas:

For GCSE and A level:

- Clarity of mark scheme instructions
- Clarity of marking principles
- Level of indicative content and terminology to guide markers as to the range of responses likely to be worthy of credit
- Whether mark schemes appropriately and fairly credit candidates for what they know, understand and can do
- For points-based mark schemes, whether there are a similar number of possible creditworthy points to the number of marks available
- Clarity of mark schemes in relation to the qualities of candidates' work that would attract higher marks
- Whether mark schemes reward only 'perfect' answers for top marks
- Whether mark schemes may advantage native or non-native speakers in any way.

A level only:

- Comparability of writing tasks both where there is a choice between tasks within a paper and to the other questions/tasks on the paper
- Comparability of the different speaking stimuli.

Subject experts completed a home-based analysis of the assessment materials, making qualitative judgements in relation to each of the areas listed above, first for A level and then for GCSE. The findings for each qualification and language were drawn together by the lead reviewer in preparation for a meeting where the main findings were identified, including any common or differing findings across the languages.

The findings of the expert review are presented in Appendices A to D for each exam board and will be referred to as required in the analysis below. Where possible, the qualitative findings are further investigated using the available quantitative data. Recommendations for further quantitative investigation of the qualitative findings are made as appropriate.

2.2 Quantitative analysis

To support this investigation, all exam boards provided the following data:

- Candidates' item-level marks for all units sat from the January 2011 series up to and including June 2013 (where available[1])
- Candidates' certification data and historical unit-level data for candidates certificating to AS or A level in summer 2013
- Designation of items to assessment objectives and skills (reading, writing and so on).

These data were used to further explore the qualitative findings of the expert reviewers in addition to supporting a wider technical review of the functioning and quality of the assessments. A summary of the data used for this analysis is provided in Table 2.1.

[1] Some exam boards do not hold mark data for spoken language assessments at a level of detail finer than the overall unit and therefore item-level data are not available. Also, optional routes are available through many of the assessments. However, practice varies in terms of whether this has been provided for each optional route or whether the data from optional routes have been aggregated.

Table 2.1: Summary of the data supplied to support quantitative analysis of A level MFL assessment functioning

	Language	Level	Summer 2013 certification data		Summer 2013 A level certificating candidates with item-level data from the series indicated						
			Entry code	Candidates in file	Unit codes	Jan 2011	Jun 2011	Jan 2012	Jun 2012	Jan 2013	Jun 2013
AQA	French	AS	1651	9,150	FREN1	2	24	6	527	169	7,523
					FREN2	-	-	-	-	-	-
		A level	2651	5,750	FREN1	6	72	74	2,311	846	1,510
					FREN2[2]	-	-	-	-	-	-
					FREN3	-	1	-	16	-	4,746
					FREN4	-	-	-	-	-	-
	German	AS	1661	3,271	GERM1	-	5	3	124	94	2,810
					GERM2	-	-	-	-	-	-
		A level	2661	1,948	GERM1	2	26	33	811	291	543
					GERM2	-	-	-	-	-	-
					GERM3	-	-	-	31	-	1,675
					GERM4	-	-	-	-	-	-
	Spanish	AS	1696	6,494	SPAN1	-	8	3	306	102	5,250
					SPAN2	-	-	-	-	-	-
		A level	2696	3,862	SPAN1	7	46	45	1,308	559	1,139
					SPAN2	-	-	-	-	-	-

[2] AQA units 2 and 4 for all languages have two options as outlined in section 3.1. The optional units are denoted by a T or V indicating that performances are marked by either a teacher or visiting examiner, respectively. The data from these optional units are not considered separately in this report and therefore, for the purposes of brevity, these units are referred to as one by using the logical numbering scheme presented here. For example, FRE2T and FRE2V are collectively referred to as to as FREN2.

					SPAN3	-	2	-	38	-	3,064
					SPAN4	-	-	-	-	-	-
OCR	French	AS	H075	1,455	F701	-	-	-	-	-	-
					F702	-	2	1	81	26	1,340
		A level	H475	993	F701	-	-	-	-	-	-
					F702	-	7	7	444	135	382
					F703	-	-	-	-	-	-
					F704	-	1	-	12	26	948
	German	AS	H076	699	F711	-	-	-	-	-	-
					F712	-	-	-	42	17	636
		A level	H476	438	F711	-	-	-	-	-	-
					F712	-	5	1	171	96	160
					F713	-	-	-	-	-	-
					F714	-	-	-	5	10	422
	Spanish	AS	H077	970	F721	-	-	-	-	-	-
					F722	-	1	1	63	19	878
		A level	H477	629	F721	-	-	-	-	-	-
					F722	-	10	13	254	99	248
					F723	-		-	-	-	-
					F724	-		1	10	3	604
Pearson	French	AS	8FR01	3,275	6FR01	3	12	8	327	147	3,434
					6FR02	1	15	3	274	80	3,494
		A level	9FR01	2,192	6FR01	7	91	50	1,677	400	432
					6FR02	4	60	21	1,577	305	624
					6FR03	-	5	-	103	-	2,564
					6FR04	-	4	-	52	-	2,455
	German	AS	8GN01	1,643	6GN01	-	10	11	194	78	1,618
					6GN02	-	6	11	100	63	1.688

		A level	9GN01	957	6GN01	5	54	95	735	141	220
					6GN02	6	34	43	504	274	358
					6GN03	-	1	-	57	-	1,193
					6GN04	-	5	-	48	-	1,143
	Spanish	AS	8SP01	2,444	6SP01	-	18	4	224	161	2,472
					6SP02	-	8	-	140	47	2,678
		A level	9SP01	1,552	6SP01	6	100	246	1,297	503	490
					6SP02	3	73	57	1,136	476	852
					6SP03	-	7	-	118	-	2,539
					6SP04	-	5	-	51	-	2,444
WJEC	French	AS	2191	3,175	FN1	-	5	-	303	-	2,814
					FN2	-	2	2	117	146	2,873
		A level	3191	1,996	FN1	-	49	-	1,515	-	343
					FN2	1	23	25	781	474	617
					FN3	-	-	-	34	-	1,915
					FN4	-	-	-	26	-	1,951
	German	AS	2221	1,368	GN1	-	3	1	99	51	1,250
					GN2	-	1	-	64	-	1,240
		A level	3221	836	GN1	-	15	-	549	-	249
					GN2	-	6	16	363	144	292
					GN3	-	-	-	15	-	815
					GN4	-	-	-	8	-	821
	Spanish	AS	2361	2,082	SN1	-	1	-	165	-	1,885
					SN2	-	1	1	89	76	1,893
		A level	3361	1,290	SN1	-	13	-	819	-	375
					SN2	-	6	13	518	293	400
					SN3	-	-	-	12	-	1,263
					SN4	-	-	-	11	-	1,268

2.3 Structure of the report

The issues raised from the expert review process and subsequent quantitative exploration are presented for each exam board in turn (sections 3 to 6). This is followed by discussion of overarching issues that result from the analysis that may affect all of the exam boards considered (sections 7 to 9). While a number of recommendations are made throughout the report, the findings and recommendations are summarised in section 10.

3 AQA

3.1 Assessment structure

The assessment structure for the current A level MFL specifications offered by AQA is summarised in Table 3.1.

Table 3.1: AQA A level MFL assessment framework

Level	Unit code	Mode of assessment	Intended weight within A level	Assessment objectives	Max raw mark
AS	FREN1 GERM1 SPAN1	Written examination	35%	AO1 = 11% AO2 = 16% AO3 = 8%	110
	FRE2T/V GER2T/V SPA2T/V	Speaking assessment (T = teacher marked, V = marked by visiting examiner)	15%	AO1 = 7% AO2 = 3% AO3 = 5%	50
A2	FREN3 GERM3 SPAN3	Written examination	35%	AO1 = 8% AO2 = 19% AO3 = 8%	110
	FRE4T/V GER4T/V SPA4T/V	Speaking assessment (T = teacher marked, V = marked by visiting examiner)	15%	AO1 = 7% AO2 = 4% AO3 = 4%	50

3.2 Subject expert scrutiny

A consolidated version of the findings from the qualitative review of AQA's A level assessment materials is provided in Appendix A. Contained in the appendix are references to further analysis provided in this paper or recommendations for further analysis/consideration. The views of the subject experts suitable for quantitative exploration using the available data are:

> 2a. For questions 1c and 1d of the Spanish unit 1 exam, the principles behind the classification of certain responses to individual items as incorrect are not clear.
>
> 4a. For all languages, the design of the mark scheme for the section B writing task of units 1 and 3 was questionable. The marks awarded for quality of language (represented by three mark grids for range of vocabulary, range of structures and accuracy) cannot be more than one band higher than the band awarded for content. This means that there is the potential for candidates' marks to be reduced three times.
>
> 5a. For questions 1g and 4c of the Spanish unit 1 exam, it is not clear whether all the information in the mark scheme boxes is required or whether these are alternative answers.
>
> 7a. There are some instances in the mark schemes (for unit 4) where the top mark bands appear to set very high performance expectations (see Appendix A for full finding).

Quantitative exploration of findings 2a and 5a is provided as part of section 3.6 of this report, finding 4a is considered in section 3.3, and finding 7a is considered in section 3.8.

3.3 Optional writing tasks

As part of the scrutiny of assessment materials, subject experts raised concerns regarding the approach taken in the award of marks for section B of the AQA written assessments (units 1 and 3) (see Appendix A, finding 4a). This section of the assessments requires candidates to produce an extended written response to one of a number of optional prompts across four or five topic areas. Twenty marks are available for content at AS level and 25 marks are available for content at A2. At both levels, 15 marks are available for the quality of language. These are allocated 5 marks for the range of vocabulary, 5 marks for accuracy and, for unit 1, 5 marks for the range of structures, and, for unit 3, 5 marks for complexity of language. This same mark structure is used for all three languages.

Candidates' responses are marked for content using a banded, levels-of-response mark scheme. Candidates are then awarded a mark for each of the three quality of language areas. However, restrictions are applied to these marks dependent on the

mark awarded for content. Instructions to examiners for the award of the quality of language marks differ at AS and A2 level and include the following statements:

> **AS (FREN1/GERM1/SPAN1)** – It should be noted that the marks awarded for each of range of vocabulary, range of structures and accuracy cannot be more than one band higher than the band awarded for content.
>
> **A2 (FREN3/GERM3/SPAN3)** – It should be noted that the marks awarded for each of range of vocabulary, complexity of language and accuracy cannot be in a higher band than the band awarded for content.

The validity of this approach is questionable for four reasons:

1. The requirement to demonstrate an understanding of the historical, economic or cultural areas required for the award of the content marks does not feature in the assessment objectives. This issue is common to all exam boards and is therefore considered further in section 8.

2. The 'cap' on quality of language marks does not appear to have a sound basis. This approach is contrary to the compensatory approach applied across general qualifications and appears particularly inappropriate given the disparate nature of the content and quality of language skills being assessed.[3]

3. Irrespective of points 1 and 2 above, the difference in the detail of the marking rules between AS and A2 appears to be specious.

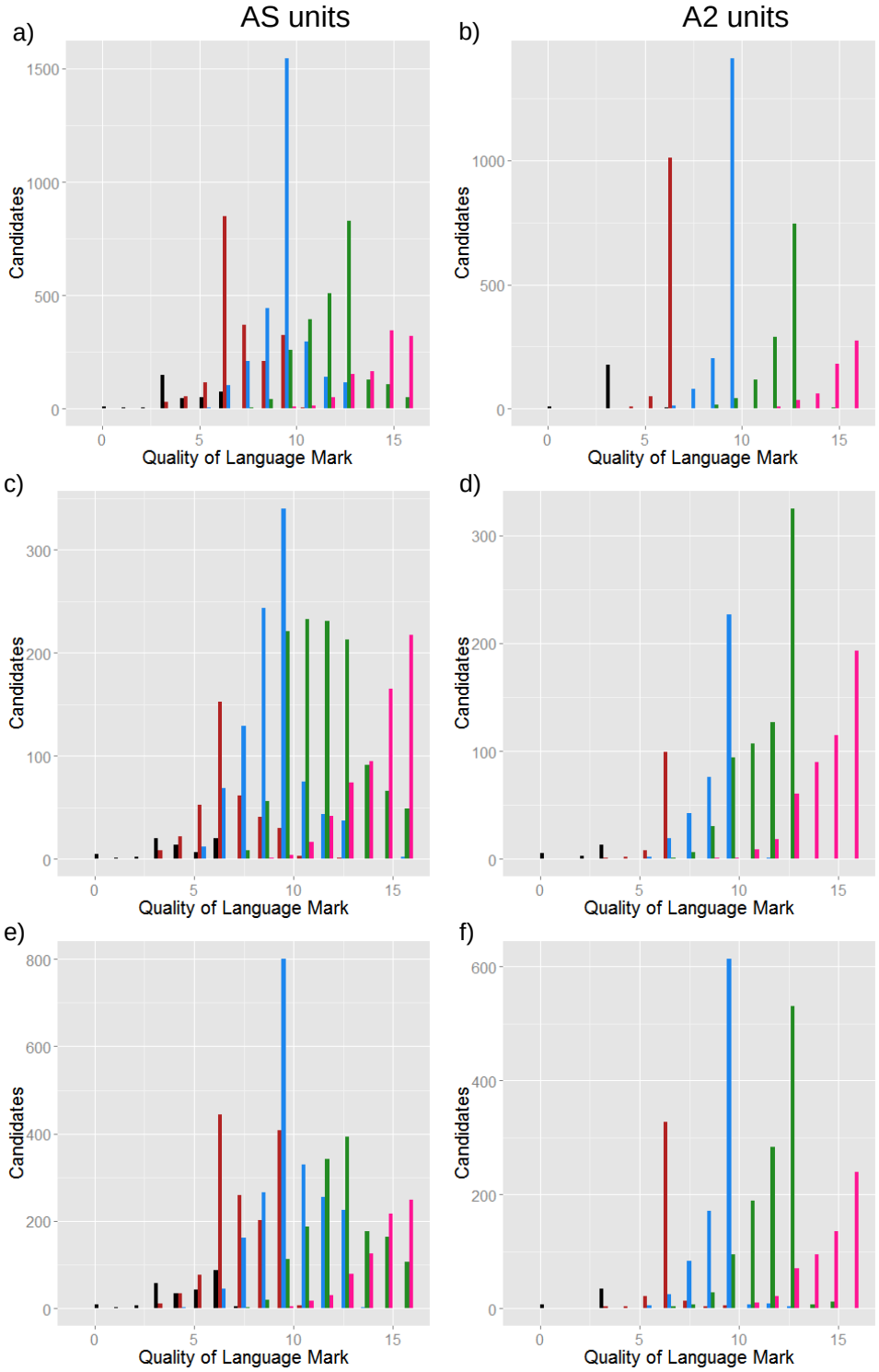
4. Any erroneous variation in the marking of content has the potential to be compounded by the permitted quality of language marks being restricted.

To investigate the impact of applying the cap on quality of language marks, the distribution of quality of language marks (summed across all three quality of language attributes) were examined. Figure 3.1 shows the quality of language mark distributions subdivided by the mark band within which candidates' content marks were awarded. This is provided for AS and A2 written units for all three languages for the summer 2013 assessments.[4]

[3] It should be noted that subsequent analysis indicates a very strong (> 0.9) correlation between content and quality of language marks. However, the strength of this relationship is argued to be artificial due to the marking restrictions applied.

[4] The data sets provided do not differentiate between the different optional items available to candidates and therefore this analysis cannot be disaggregated between routes.

Figure 3.1: Distribution of candidates' quality of language marks subdivided by allocation to content band in a) FREN1, b) FREN3, c) GERM1, d) GERM3, e) SPAN1 and f) SPAN3 for the AQA summer 2013 assessments[5]



[5] Content band key: black = 1, red = 2, blue = 3, green = 4, pink = 5. Labels on the *x* axes are located to the left of the leftmost bar in any grouping

Notable features of these distributions are as follows:

1. The cap on quality of language marks appears to have had a significant impact on the shape of the mark distribution in all A2 units with a large proportion of candidates receiving the maximum marks permitted for quality of language. If this rule were not in place, the shapes of the distributions for quality of language are likely to have extended beyond this capped mark. The clustering of candidates at the top of the permitted number of marks for the separate quality of language skills is demonstrated in Tables 3.2 to 3.4 where it is shown that there is a very high proportion of candidates being awarded at least two out of their three quality of language marks at the maximum level they are permitted.

2. The cap appears to have impacted on the shape of the mark distribution in the AS units. In a similar way to that observed with the A2 units, the distributions are likely to have been truncated relative to their shape had this cap not been in place. Further, seemingly anomalous peaks in these distributions occur three marks lower than the cap on quality of language marks. While this may be coincidental, this strongly suggests that examiners may have been inappropriately applying the rule from the A2 mark scheme when marking responses on the AS units – capping the quality of language marks to the content band rather than to the band above. This may therefore have led to a reduced number of marks for some candidates.

3. There are a small number of instances where candidates have been awarded marks that do not conform to these rules. The data behind Figure 3.1 show, for the A2 units, 0.28 per cent of marks in French, 0.04 per cent of marks in German and 1.19 per cent of marks in Spanish were awarded higher than permitted by the mark scheme based on the marks candidates were awarded for content.

The decision to cap quality of language marks according to the content written by a candidate was considered appropriate by those subject experts responsible for the design of the mark scheme. However, the evidence suggests that the rule has unintended consequences that in practice are likely to impact on the rank order of candidates and, therefore, may have negative consequences on the validity of the mark distribution.

Table 3.2: Percentage of candidates achieving at least 2 out of 3 quality of language marks in the section B of FREN3 in June 2013[6]

Content band		Percentage of candidates in the content band achieving at least 2 out of 3 marks in the indicated quality of language mark band					No. of candidates in content band
Band number	Mark range	QoL band 1	QoL band 2	QoL band 3	QoL band 4	QoL band 5	
1	0–5	97.4	2.6	0.0	0.0	0.0	189
2	6–10	0.8	98.9	0.3	0.0	0.0	1,071
3	11–15	0.0	5.3	94.6	0.1	0.0	1,709
4	16–20	0.0	0.2	13.5	86.0	0.3	1,204
5	21–25	0.0	0.0	0.5	16.9	82.5	555

Table 3.3: Percentage of candidates achieving at least 2 out of 3 quality of language marks in the Section B of GERM3 in June 2013

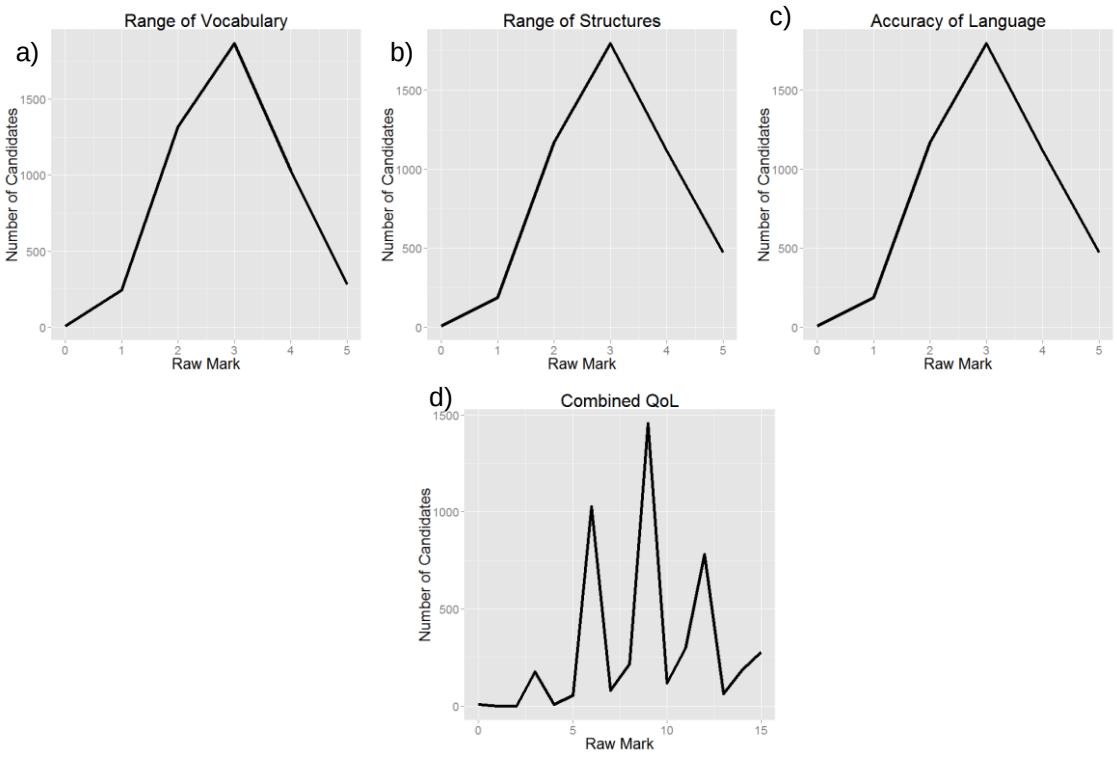
Content band		Percentage of candidates in the content band achieving at least 2 out of 3 marks in the indicated quality of language mark band					No. of candidates in content band
Band number	Mark range	QoL band 1	QoL band 2	QoL band 3	QoL band 4	QoL band 5	
1	0–5	100.0	0.0	0.0	0.0	0.0	21
2	6–10	2.7	97.3	0.0	0.0	0.0	110
3	11–15	0.0	16.0	83.7	0.3	0.0	367
4	16–20	0.0	0.9	32.6	66.5	0.0	681
5	21–25	0.0	0.0	2.3	30.9	66.7	475

[6] QoL = quality of language (Tables 3.2 to 3.4).

Table 3.4: Percentage of candidates achieving at least 2 out of 3 quality of language marks in the section B of SPAN3 in June 2013

Content band		Percentage of candidates in the content band achieving at least 2 out of 3 marks in the indicated quality of language mark band					No. of candidates in content band
Band number	Mark range	QoL band 1	QoL band 2	QoL band 3	QoL band 4	QoL band 5	
1	0–5	97.7	2.3	0.0	0.0	0.0	43
2	6–10	1.9	95.8	2.4	0.0	0.0	377
3	11–15	0.0	11.9	86.8	1.3	0.0	911
4	16–20	0.0	0.7	26.1	72.1	1.1	1142
5	21–25	0.0	0.0	2.2	29.3	68.5	553

Figure 3.2: Mark distributions for the separate quality of language skills (a to c) for FREN3 from June 2013 and the combined effect (d)



To investigate the impact on discrimination between candidates, the mark distributions for the separate quality of language areas are shown in Figures 3.2a to 3.2c using unit FREN3 as an example. These mark distributions appear to discriminate well between candidates. However, when combined across the quality of language skills (Figure 3.2d), the reduced discrimination between candidates is observed, with candidates clustering on marks 3, 6, 9, 12 and 15. It may be argued that a mark distribution such as that presented in Figure 3.2d may legitimately occur when combining marks across three highly correlated items. However, when this evidence is combined with that presented above, it is highly likely that the reduced discrimination and the atypical distribution of marks have been imposed by the quality of language marking rules applied.

3.4 Impact of marking rules on grading

The presence of the cap on quality of language marks is likely to have led to candidates being awarded fewer marks than would have been the case had this rule not been in place. However, this would not necessarily reduce the proportion of candidates receiving higher grades. The use of statistical evidence during grade boundary setting would protect against this. While the appropriate placement of grade boundaries can address any such effect, it cannot, however, address any issues with candidate rank order or reduced discrimination.

3.5 Appropriateness of demand

Shown in Figures 3.3, 3.4 and 3.5 are the item-level mark distributions for the written units FREN3, GERM3 and SPAN3, respectively, from the June 2013 exam series for A level certificating candidates only. These distributions are also summarised by the item facility indices shown in Table 3.5.

Across all languages, many of these items have a high facility index implying low demand for these candidates and the mark distributions exhibit negative skew likely leading to poor discrimination between candidates at the top of the ability range. No item across any of the assessments has a facility index lower than 0.5. The consequences of this prevalence of relatively low-demand items on the unit-level mark distributions are shown Figure 3.6, with the descriptive statistics provided in Table 3.6. It is clear from these distributions that a significant proportion of the lower end of the mark distribution is not used to discriminate between candidates. The relatively long mark scales available on these assessments does, however, provide some protection against the reduced discrimination of these assessments caused by the issues outlined above.

Figure 3.3: Item-level mark distributions for FREN3 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

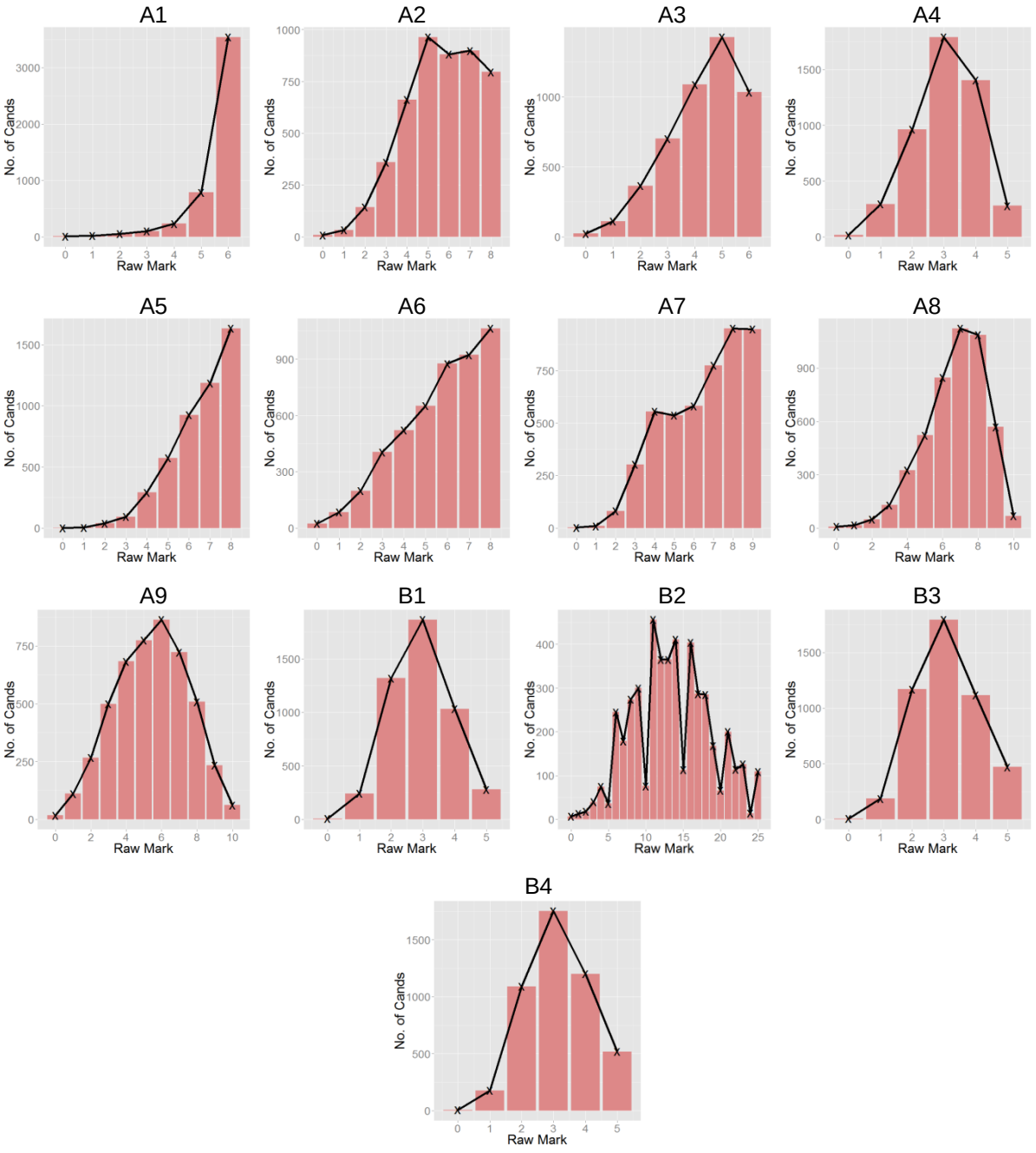


Figure 3.4: Item-level mark distributions for GERM3 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

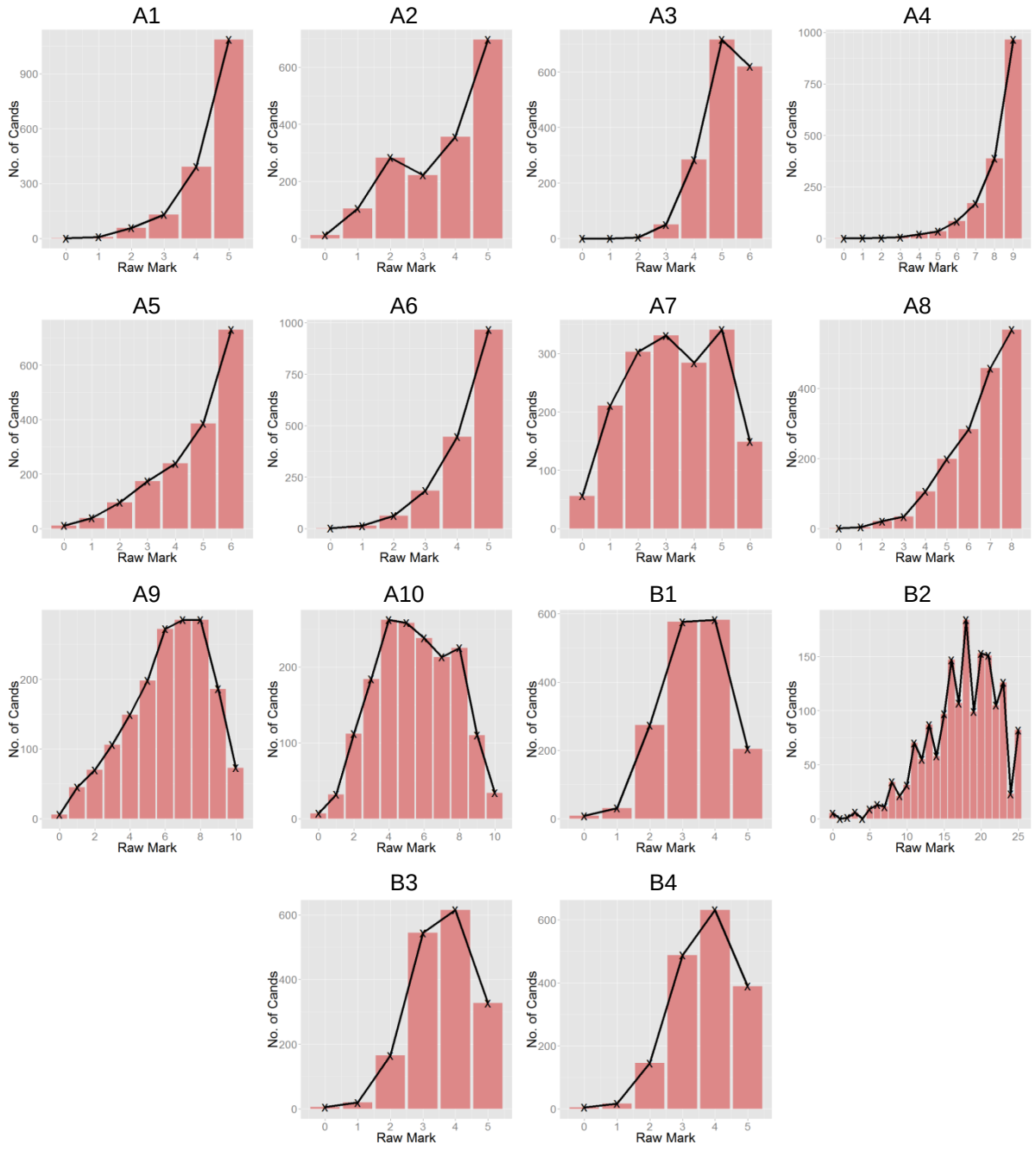


Figure 3.5: Item-level mark distributions for SPAN3 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

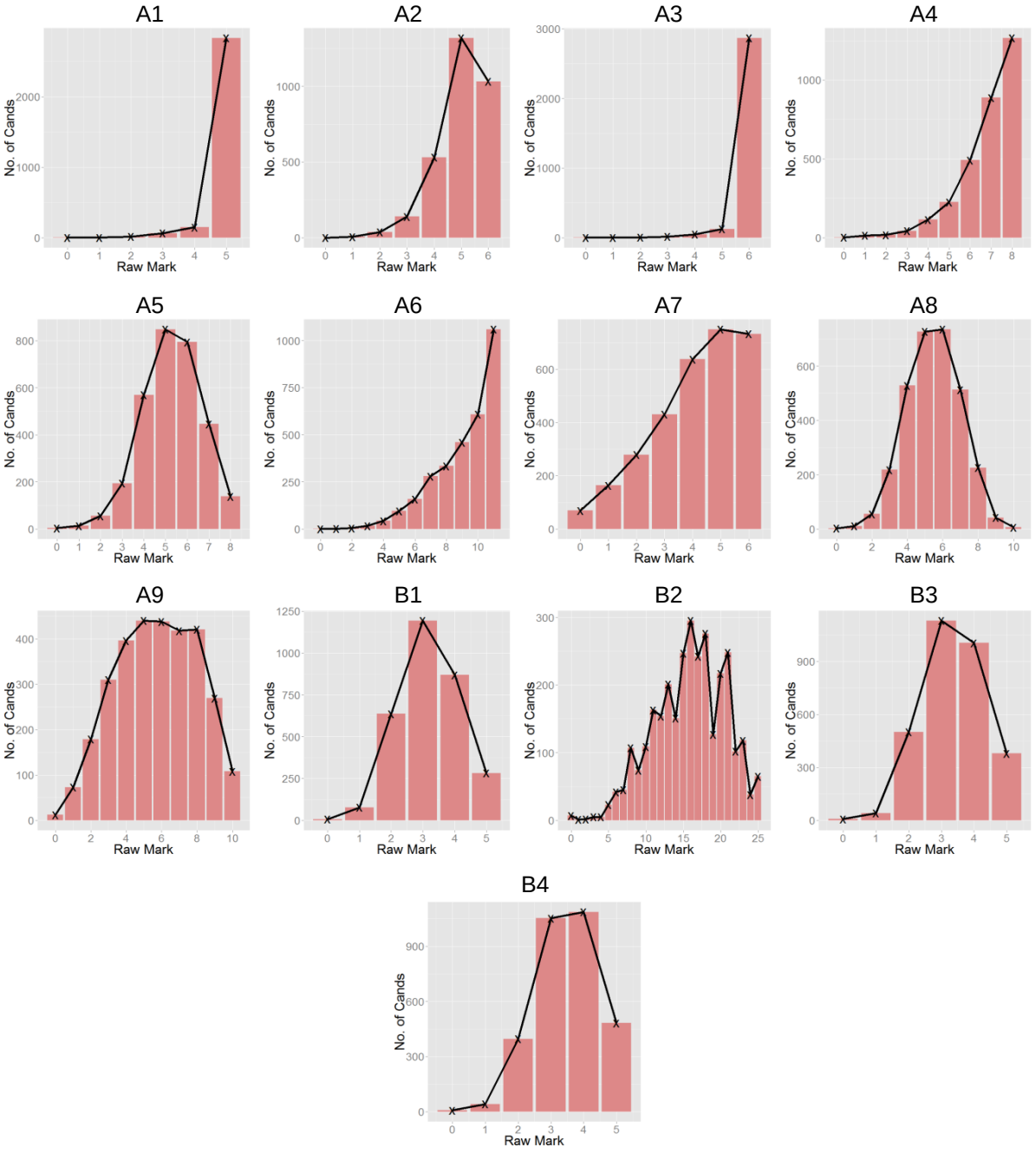
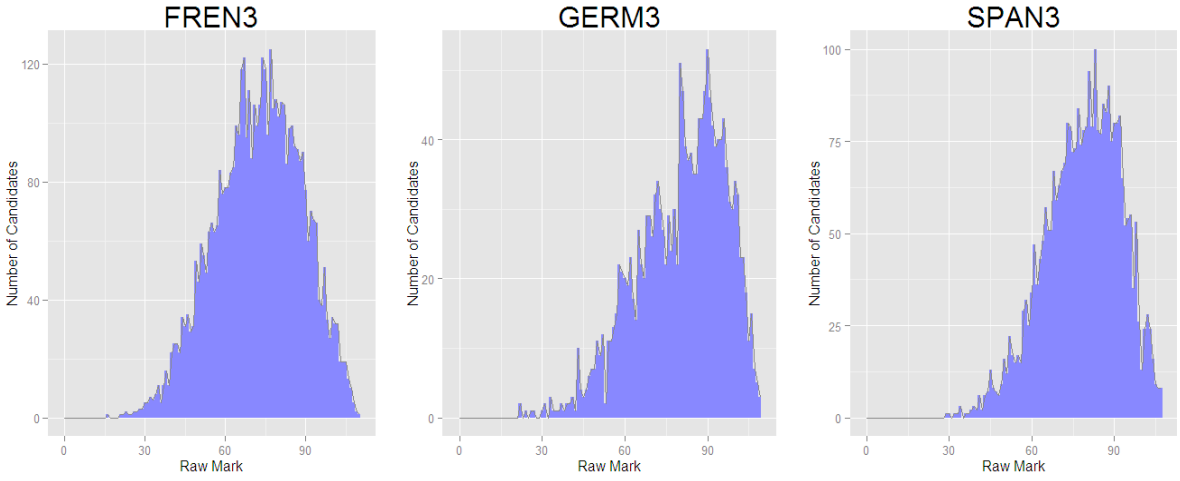


Table 3.5: Item facility indices for FREN3, GERM3 and SPAN3 for certificating candidates in June 2013

Item		Item facility index		
		FREN3	GERM3	SPAN3
Section A	1	0.93	0.90	0.98
	2	0.71	0.75	0.84
	3	0.72	0.86	0.98
	4	0.62	0.91	0.86
	5	0.82	0.80	0.66
	6	0.72	0.87	0.84
	7	0.73	0.55	0.69
	8	0.67	0.82	0.55
	9	0.54	0.62	0.57
	10	-	0.55	-
Section B[7]	1	0.59	0.68	0.64
	2	0.54	0.70	0.64
	3	0.62	0.73	0.68
	4	0.63	0.75	0.70

Figure 3.6: Unit-level mark distributions for the AQA A2 written units for A level certificating candidates in June 2013



[7] Note that section B is composed of ten optional items. However, for the purposes of this analysis no distinction is made between the different options. The items in section B, as labelled in this table, relate to 1 = range of vocabulary (quality of language), 2 = content, 3 = complexity of language (quality of language), 4 = accuracy (quality of language).

Table 3.6: Descriptive statistics for units FREN3, GERM3 and SPAN3 from June 2013 for certificating A level candidates only

Unit	Mean	Standard deviation	Skewness[8]
FREN3	72.5 (65.9%)	16.0 (14.5%)	-0.22
GERM3	80.5 (73.2%)	16.0 (14.5%)	-0.63
SPAN3	78.5 (71.4%)	13.7 (12.5%)	-0.42

To evaluate the quality of the measurement provided by these assessments, the partial credit model was fitted to the item-level data sets for the A2 units.[9] This enabled assessment of the item and test information available across the ability range. Shown in Figure 3.7 are the level information functions, respectively for the three units. This function indicates where on the latent scale candidates ability is being most effectively measured. Superimposed on these distributions are the unit-level grade boundaries[10] transformed from raw mark space to the latent trait scale. These allow required standard and the ability of candidates to be compared against the effectiveness of the assessment to measure ability. To provide an indication of model fit, the expected and empirical item-level category probability curves and item characteristic curves are provided in Appendices F, G and H for units FREN3, GERM3 and SPAN3, respectively.[11]

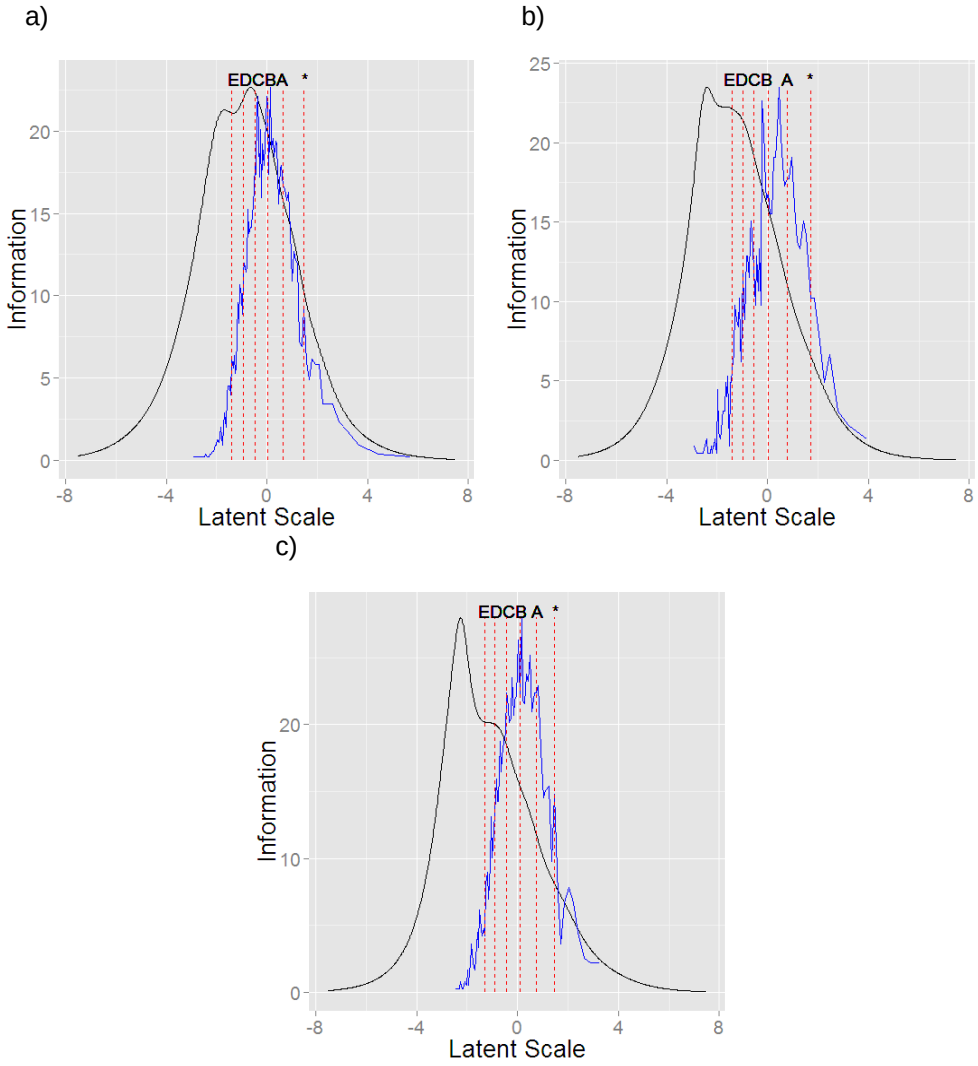
[8] Defined as the third standardised moment of the raw mark distribution.

[9] Fitting this model to data containing items with tariffs as high as those contained in AQA written assessments is pushing its applicability to its limits. However, given the use to which the model is being put in this context, this is not considered overly problematic. To demonstrate the impact of removing the higher tariff items from the analysis, Figure 3.7 is replicated in Appendix O for section A only of the A2 written exams.

[10] In addition to the unit-level A* UMS conversion point.

[11] As the joint maximum likelihood approach was used for parameter estimation and therefore the fit was conditioned on candidates, expected and observed scores being identical, the unit-level observed and expected mark distributions were identical.

Figure 3.7: Test information functions (black) for a) FREN3, b) GERM3 and c) SPAN3 from June 2013. Superimposed are the unit-level grade boundaries (dotted) and the distribution of candidate person parameters relative to the information functions (blue).



This analysis is helpful because, while the facility indices provide an indication of the demand for the candidates sitting the assessment, it does not provide an indication of the appropriateness of the demand relative to the required standard (represented by the grade boundaries). This analysis also looks at how effective the assessments are at collecting information around the grade boundaries.

Figure 3.7 shows that there is a misalignment between the information function and the grade boundary locations. The peaks in the distributions show that FREN3 is most effective at measuring candidates achieving a grade D, whereas the ability of candidates achieving below the E grade boundary is most effectively measured in GERM3 and SPAN3. In other words, insufficient information is gathered around the higher-grade boundaries given the length of the assessments. This is most notable

for German and Spanish. This mismatch will have impacted on the quality of measurement for the most able candidates, therefore providing a less reliable measure than desirable.

3.6 Exploration of item-level qualitative findings

The findings of the expert reviewers were investigated further. Specifically, findings 2a: that the definition of incorrect responses was not clear for questions 1c and 1d of SPAN1;[12] and 5a: that it was not clear whether all the information in the mark scheme boxes is required or whether these are alternative answers for questions 1g and 4c[13] of SPAN1. The item summary statistics for SPAN1 in June 2013 are provided in Table 3.7.

Table 3.7: Item summary statistics for SPAN1 in June 2013

Item reference	Facility index	Discrimination index[14]	No. of candidates
A1	0.66	0.74	6,008
A2	0.73	0.71	
A3	0.72	0.69	
A4	0.73	0.79	
A5	0.79	0.57	
A6	0.60	0.79	
A7	0.69	0.67	
A8	0.61	0.79	
A9	0.46	0.78	
B1	0.61	0.78	
B2	0.56	0.79	
B3	0.65	0.80	
B4	0.69	0.78	

There are no notable differences between the item summary statistics for items A1 and A4 that were the subject of these qualitative findings and other items. However, as the available data are reported at the overall question level rather than for the sub-items making up the questions, the magnitude of any effect may be reduced. Therefore, despite no quantitative evidence being available to demonstrably support

[12] Sub-items of item A1.

[13] Sub-item of item A4.

[14] Defined as the correlation between marks on the items and marks on the rest of the assessment excluding that item.

the qualitative finding, it is recommended that the opinion of the expert reviewers be considered when developing similar items in the future.

3.7 Rank ordering of candidates

The analysis in section 3.5 provides consideration of the demand of the items, but sheds no light on the appropriateness of the rank order of candidates within those distributions. The best operationally accessible indicator of the validity of mark distributions is the correlation of candidates' marks with other indicators of performance deemed to be equivalent. Marks for an assessment or subset of items that correlate weakly with a measure deemed equivalent may indicate that issues exist with the rank order of candidates. To this end, the relationship between candidates' marks on listening, reading and writing on the AS and A2 assessments were examined. The designation of items to these skills for the June 2012 AS units and June 2013 A2 units[15] are as quoted in Table 3.8, with a key to the abbreviations provided in Appendix E. All reading and listening items at both AS and A2 are appropriate for objective marking as they are either matching or multiple-choice questions. The writing items are those discussed in section 3.3 that comprise marks for both content and quality of language. Table 3.9 shows the intra-skill correlation coefficients for the written assessments for candidates who sat the AS written assessment in summer 2012, the A2 written assessment in summer 2013 and certificated at A level in summer 2013. The corresponding scatter plots are shown in Figure 3.8.

The correlation coefficients for reading and listening are reasonable. However, it is clear from consideration of Figure 3.8 that these data contain a significant ceiling effect on both axes. These effects can impact significantly on the value of correlation coefficients. Groups of candidates whose marks approach the maximum mark on just one axis may lead to the correlation being reduced. However, groups of candidates who approach the maximum mark on both axes may increase the correlation. To reduce these competing effects (both of which act to distort consideration of the underlying relationship), candidates scoring within two marks of the maximum mark at either AS or A2 the intra-skill correlations were removed and the correlations recalculated. These modified correlations are reported in Table 3.10. The lower value of these modified correlations compared to those presented in Table 3.9 suggest that the net results of the ceiling effect is to increase the correlations.[16] The larger impact of this recalculation on listening and reading compared to writing shows that the ceiling effect on this calculation was greater for these skills.

[15] The units/series in which candidates certificating at A level in June 2013 are most likely to have taken.

[16] A small additional component contributing to this reduction in correlation is likely to arise from the reduction in spread of marks by this filtering of the data.

The correlations for writing are lower than for the other two skills. This lower correlation is likely to arise from one, some or all of the following potential sources:

1. More variability in the progression of candidates in their writing skills between AS and A2 than in reading and listening

2. Lower accuracy of marking arising from the subjective nature of the application of a levels-of-response mark scheme

3. An effect on rank order due to inconsistent application of the capping rules at AS

4. An effect due to differences in optional routes taken through the assessments that is not visible from the available data sets.

Given that sources 2, 3 or 4 seem more likely (there is no evidence that we know of to support 1), this suggests a reduction in the validity of the rank order compared to that defined for listening or reading. However, it is not possible to disaggregate the degree to which this arises from a legitimate difference in professional opinion during marking and the other illegitimate sources of variation.

Table 3.8: Designation of items to skills for the AQA written assessments

Item number	June 2012	June 2013	
	FREN1 / GERM1 / SPAN1	FREN3 / SPAN3	GERM3
A1	L	L	L
A2	L	L	L
A3	L	L	L
A4	L	L	L
A5	R	R	R
A6	R	R	R
A7	R	R	R
A8	R	RW	R
A9	RW	RW	RW
A10	-	-	RW
B1	WO	WO	WO
B2	WO	WO	WO
B3	WO	WO	WO
B4	WO	WO	WO

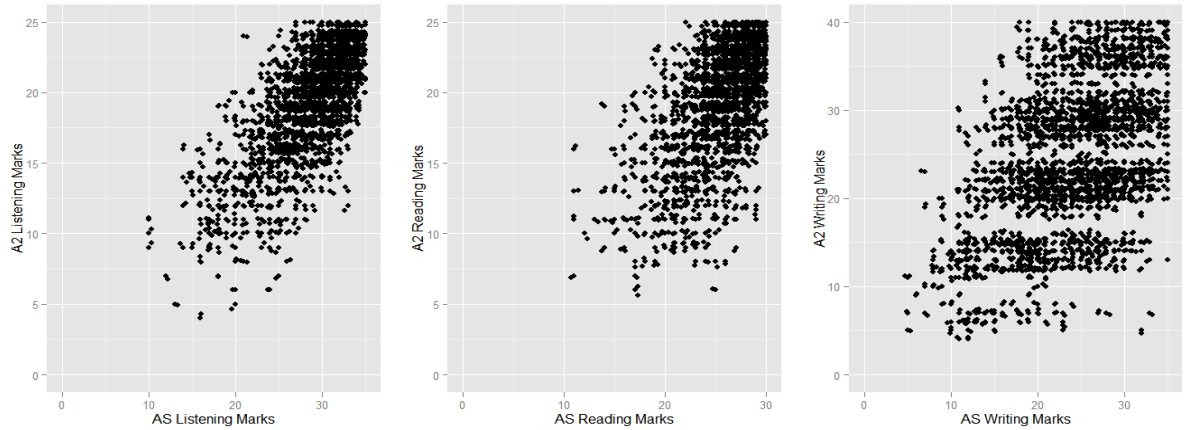
Table 3.9: Intra-skill correlations between AS and A2 levels in summer 2012 and summer 2013, respectively

	Skill	Items used for comparison		Item totals	SD – % of item(s) max	A2:AS correlation[17]	No. of candidates
French	Listening	AS	A1 A2 A3 A4	35	11.8	0.73	2,271
		A2	A1 A2 A3 A4	25	19.0		
	Reading	AS	A5 A6 A7 A8	30	10.7	0.64	2,271
		A2	A5 A6 A7	25	20.3		
	Writing	AS	B1 B2 B3 B4	35	17.8	0.51	2,267
		A2	B1 B2 B3 B4	40	22.6		
German	Listening	AS	A1 A2 A3 A4	35	11.0	0.73	791
		A2	A1 A2 A3 A4	25	19.4		
	Reading	AS	A5 A6 A7 A8	30	15.8	0.78	791
		A2	A5 A6 A7 A8	25	22.2		
	Writing	AS	B1 B2 B3 B4	35	15.5	0.59	789
		A2	B1 B2 B3 B4	40	21.7		
Spanish	Listening	AS	A1 A2 A3 A4	35	11.1	0.65	1,282
		A2	A1 A2 A3 A4	25	18.7		
	Reading	AS	A5 A6 A7 A8	30	11.9	0.74	1,282
		A2	A5 A6 A7	25	20.5		
	Writing	AS	B1 B2 B3 B4	35	17.0	0.56	1,278
		A2	B1 B2 B3 B4	40	22.5		

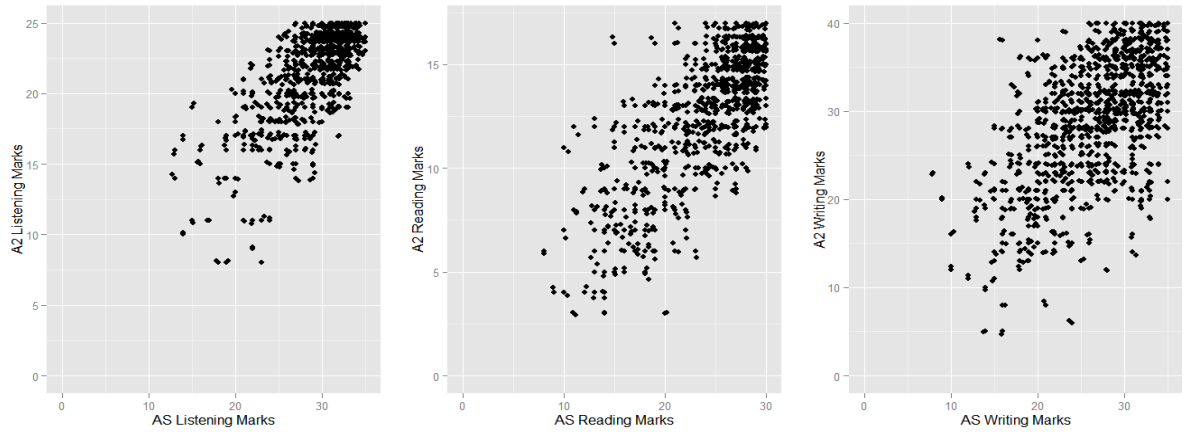
[17] The correlation coefficients are evaluated with candidates removed who scored a total of zero marks for either all AS or all A2 items identified.

Figure 3.8: Scatter plots[18] showing candidates' marks for listening, reading/writing and writing skills in French, German and Spanish at AS and A2 level for candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

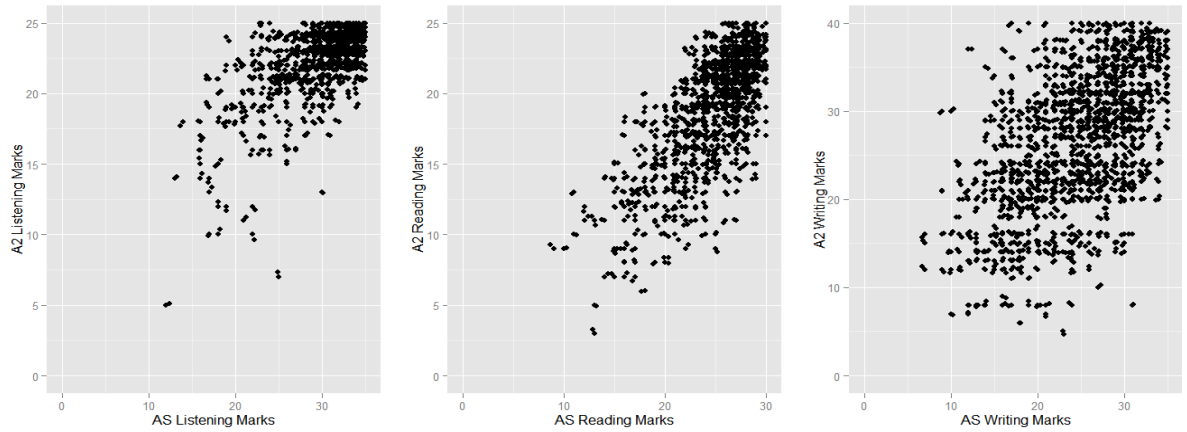
a) French



b) German



c) Spanish



[18] The data points have been jittered to avoid over-plotting to show the density of points.

Table 3.10: Intra-skill correlations between AS and A2 levels in summer 2012 and summer 2013, respectively, with the ceiling effect removed

	Skill	AS:A2 correlation	A2:AS modified correlation	No. of candidates in modified correlation
French	Listening	0.73	0.69	1,479
	Reading	0.64	0.50	1,111
	Writing	0.51	0.47	1,970
German	Listening	0.73	0.59	296
	Reading	0.78	0.70	387
	Writing	0.59	0.55	595
Spanish	Listening	0.65	0.55	308
	Reading	0.74	0.69	763
	Writing	0.56	0.51	1,101

3.8 Weighting of skills

If the functioning of items is systematically dependent on the skill they measure (for example listening, reading and so on), there may be an impact on the extent to which those skills contribute to the overall ranking of candidates compared to the intended assessment design. To investigate this, the achieved weight[19] for each of the skills was evaluated and compared with the intended weight.[20] The achieved and intended weightings of the skills are presented in Table 3.11.

Table 3.11: Achieved vs intended weight for listening, reading and writing in the AQA A2 written assessments

Unit	Skill designation	Intended weight %	Achieved weight %	Difference (% pnts)
FREN3	L	22.7	18.2	-4.5
	R	22.7	21.3	-1.4
	RW	18.2	17.6	-0.6
	W	36.4	43.0	+6.6
GERM3	L	22.7	16.1	-6.6
	R	22.7	22.7	+0.0
	RW	18.2	22.0	+3.8
	W	36.4	39.2	+2.8
SPAN3	L	22.7	11.3	-11.4
	R	22.7	22.7	+0.0
	RW	18.2	20.8	+2.6
	W	36.4	45.3	+8.9

The figures quoted in Table 3.8 confirm that the reduced discrimination of the listening items have led to this skill having a lower than designed impact on the rank order of candidates. The writing task has a systematically greater impact on the rank order than intended. This feature is concerning given the reservations regarding the validity of the rules imposed on the marking of the writing tasks as outlined in section 3.3.

It may be suggested that listening is inherently less demanding than writing and so will inevitably result in items that have higher facility indices and so are less likely to discriminate between the most able candidates. However, if an assessment is to function effectively, and performance in the individual skills contribute appropriately

[19] Calculated as: achieved weight of skill = $100 \times \frac{(\text{standard deviation of marks for the skill}) \times (\text{skill-to-total correlation})}{\text{standard deviation of total marks}}$

[20] Defined as the percentage of raw marks on the assessment assigned to that skill.

to the overall grade, then the demand of the assessment of the different skills needs to be such that the intended weightings are achieved.

3.9 Speaking assessments

Finding 7a arising from the expert review of assessment materials suggested that the expectations set by the top mark bands for the unit 4 speaking assessment were very high. In Figure 3.9 are the unit-level mark distributions for units FREN4, GERM4 and SPAN4 from June 2013 for certificating A level candidates. Unit-level descriptive statistics are in Table 3.12. These distributions are skewed considerably towards the higher marks in all three subjects. This suggests that, even though the expectations appear to be very high at the top of the mark scale, candidates are either meeting those expectations or the expectations are being considered in the assessment context of what it would be reasonable for a candidate to produce under those circumstances. Indeed, the expectations of candidates of this level may be inappropriately low, or compensation for the context when setting the marking standard may be excessive given that the lower half of all three mark distributions is largely unused, with very high mean marks and negative skew.

Figure 3.9: Unit-level raw mark distributions for FREN4, GERM4 and SPAN4 for certificating A level candidates in summer 2013 with grade boundaries superimposed

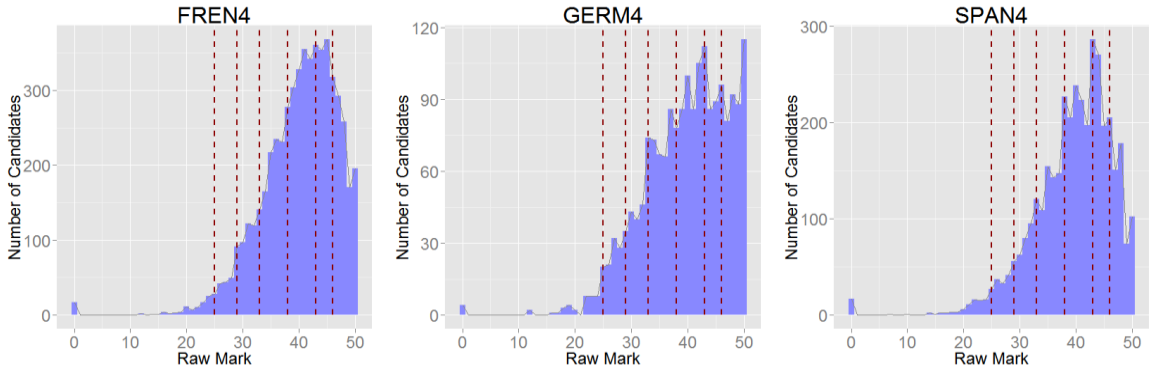


Table 3.12: Descriptive statistics for units FREN4, GERM4 and SPAN4 from June 2013 for certificating A level candidates only

Unit	Mean		Standard deviation		Skewness
FREN4	40.2	(80.4%)	6.6	(13.2%)	-1.22
GERM4	39.7	(79.4%)	7.3	(14.6%)	-0.81
SPAN4	39.5	(79.0%)	6.9	(13.8%)	-1.38

The consequences of distributions with these properties are likely to impact most strongly on the more able candidates, since Figure 3.9 shows that the distributions are truncated at the maximum mark offering no differentiation between these high-attaining candidates. Similar to the analysis performed in section 3.7, Figure 3.10 and Table 3.13 summarise the relationship between candidates' spoken language marks at AS in June 2012 (FREN2, GERM2 and SPAN2) and A2 in June 2013 (FREN4, GERM4 and SPAN4). Mindful of the challenges with maintaining and setting a marking standard for a unit of this type, consideration should be given to how the valid discrimination can be improved.

Figure 3.10: Scatter plots[21] showing candidates' marks for speaking in French, German and Spanish at AS and A2 level for candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

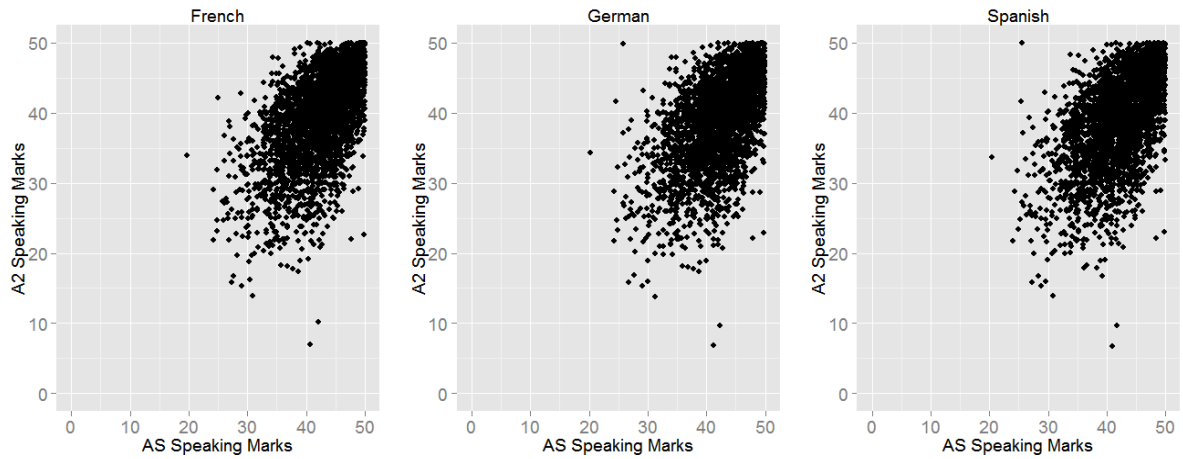


Table 3.13: Correlation and (modified correlation accounting for the ceiling effect) between candidates' marks achieved in speaking at AS level (unit 2) and A2 level (unit 4)

	AS:A2 correlation	No. of candidates in correlation	A2:AS modified correlation	No. of candidates in modified correlation
French	0.59	5,213	0.49	3,982
German	0.65	1,623	0.51	1,206
Spanish	0.58	3,371	0.49	2,662

[21] The data points have been jittered to avoid over-plotting to show the density of points.

4 OCR

4.1 Assessment structure

The assessment structure for the current A level MFL specifications offered by OCR is summarised in Table 4.1.

Table 4.1: OCR A level MFL assessment framework

Level	Unit code	Mode of assessment	Intended weight within A level	Assessment objectives	Max raw mark
AS	F701 F711 F721	Speaking tasks	15%	AO1 = 8.75% AO2 = 3.75% AO3 = 2.5%	60
	F702 F712 F722	Written examination	35%	AO1 = 8.75% AO2 = 16.25% AO3 = 10%	140
A2	F703 F713 F723	Speaking tasks	15%	AO1 = 7.5% AO2 = 2.5% AO3 = 5%	60
	F704 F714 F724	Written examination	35%	AO1 = 7.5% AO2 = 20% AO3 = 7.5%	140

4.2 Subject expert scrutiny

Two points were raised by the expert reviewers that can be directly investigated using the available data. Findings 2c and 5a in Appendix B relating to the written assessments in Spanish state:

> 2c. In unit F724, task 7, translation into English, the principle of having “night-time protest” as an acceptable answer but the reason for not allowing “night protest” is not clear, particularly considering this is a transfer of meaning exercise.
>
> 5a. In unit F722, task 3e gives two points of information in the answer box – “wide range” and “reasonably priced” – but is only worth one mark.

These findings will be investigated via the exploration of item functioning metrics in section 4.4.

4.3 Item functioning

Figures 4.1, 4.2 and 4.3 show the item-level mark distributions for the A2 written exam units for French (F704), German (F714) and Spanish (F724), respectively. Item-level descriptive statistics for the summer 2013 AS written papers are presented in Table 4.2. The equivalent statistics for the A2 units are shown in Table 4.3. The references to section A and section B for French and Spanish relate to the marks awarded to candidates for quality of language in these sections of the paper. In section A and section B, candidates are marked on their accuracy of language (maximum of 5 marks for each section). For German, this distinction is not made for section A in the supplied data sets with the quality of language marks being provided aggregated with the marks for question 2. For all languages, section C contains a single task and is assessed for relevance and points of view (10 marks), structure and analysis (15 marks), quality of language *(*accuracy) (10 marks) and quality of language (range) (10 marks) and these marks are reported aggregated in the data sets.

The item-level mark distributions show that the majority of the items from the OCR A2 written exams spread candidates across the mark distribution to an acceptable extent. Despite the reasonable spread of marks across the six written papers considered in Table 4.2, only four of the 59 (7 per cent) items have a facility index lower than 0.5, suggesting that, in general, items have reasonably low demand for the cohort.

The consequences of this item functioning on the unit-level mark distributions are shown (for the A2 units) in Figure 4.4, with the corresponding descriptive statistics in Table 4.3. These figures show mark distributions for all units that are all slightly negatively skewed, with an unused region of the mark scale at the bottom of the distribution.

Figure 4.1: Item-level mark distributions for F704 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

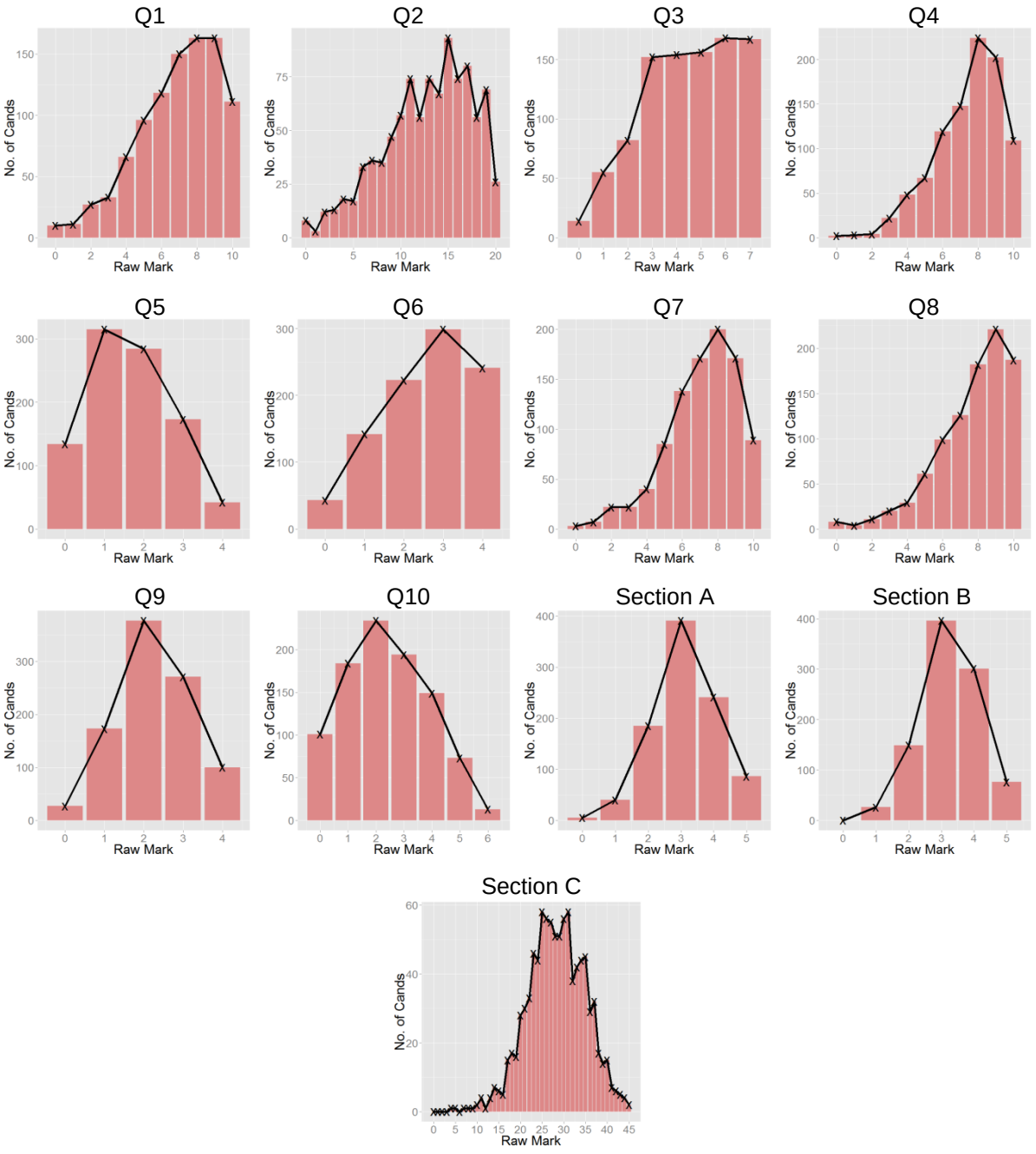


Figure 4.2: Item-level mark distributions for F714 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

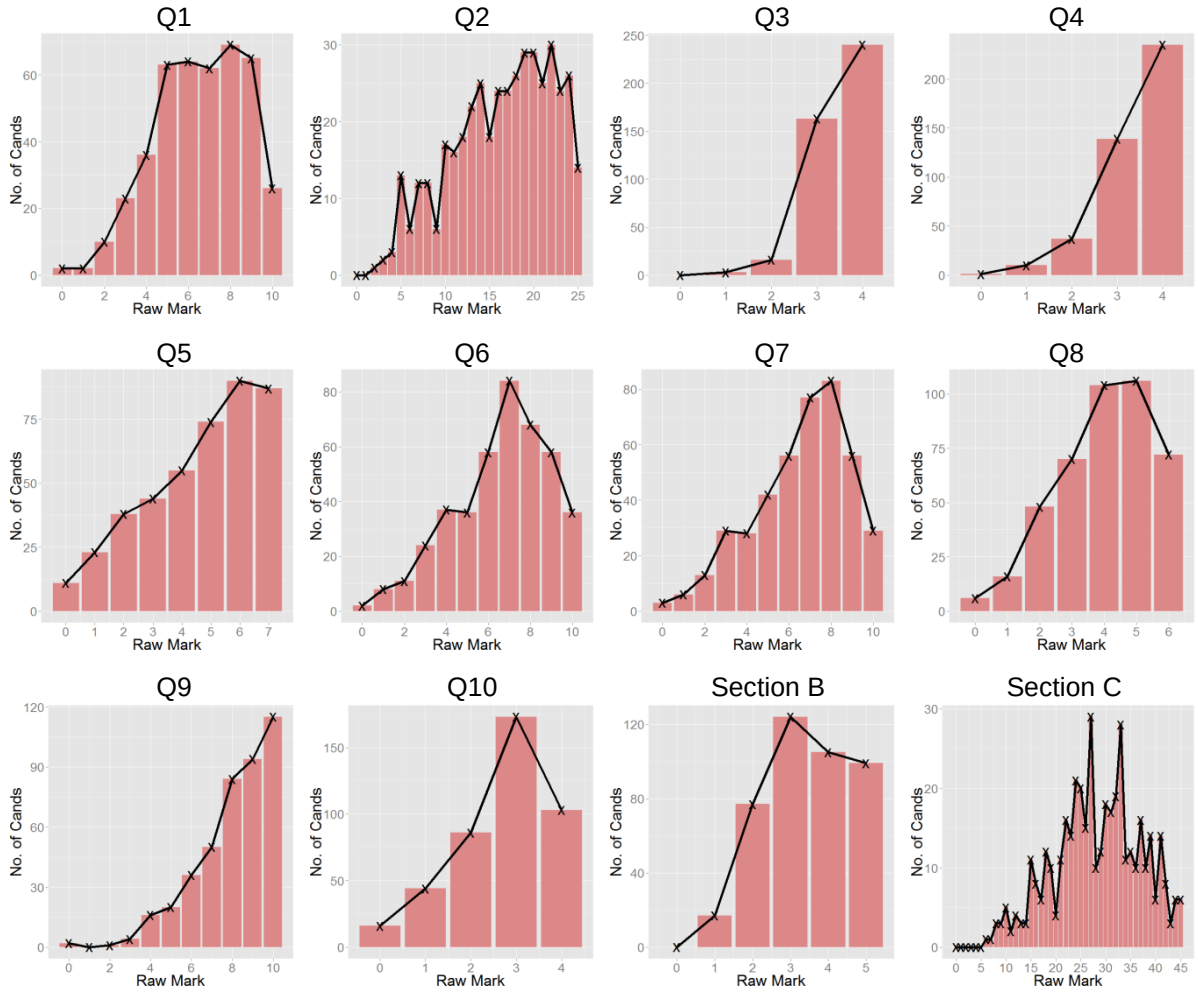


Figure 4.3: Item-level mark distributions for F724 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

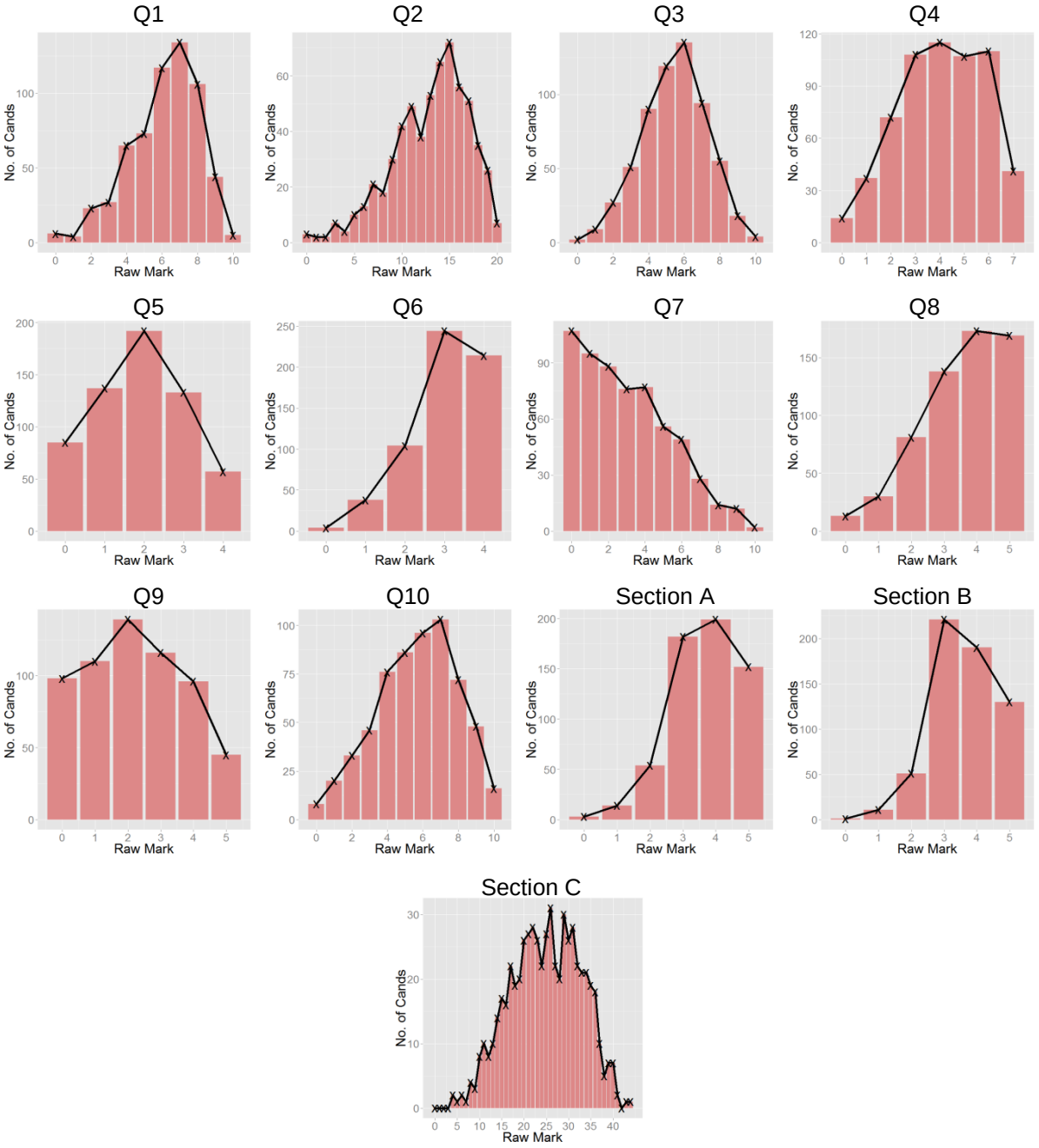


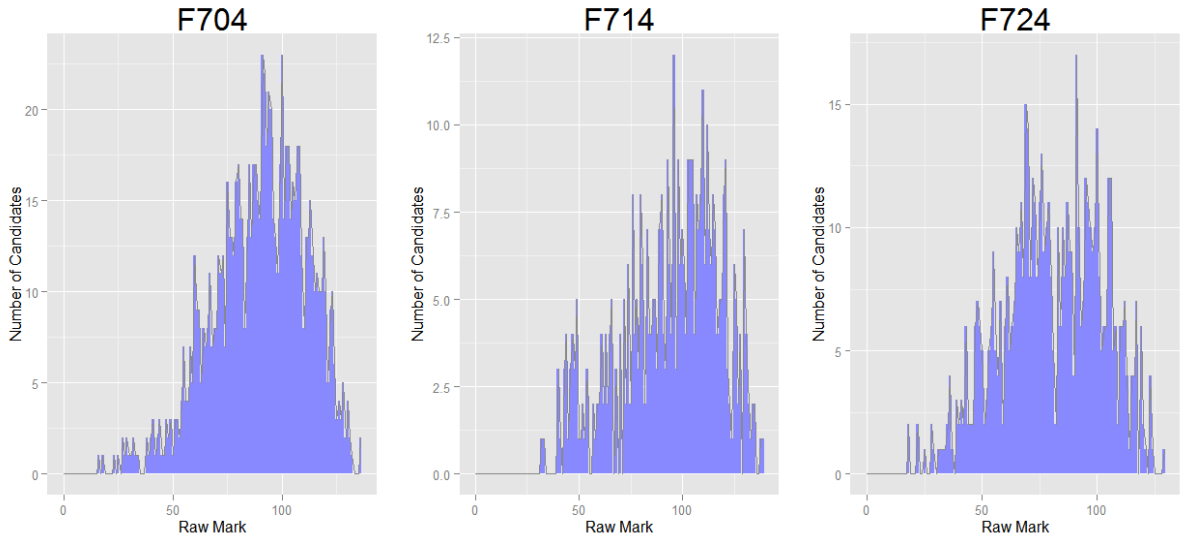
Table 4.2: Item-level descriptive statistics for the OCR written exam units

	Unit	Item reference	Facility index	Discrimination index	No. of candidates
French	F702	1	0.75	0.69	1,340
		2	0.72	0.70	
		3	0.61	0.81	
		4	0.76	0.85	
		5	0.74	0.72	
		6	0.63	0.88	
		7	0.67	0.91	
German	F712	1	0.83	0.76	636
		2	0.78	0.79	
		3	0.68	0.81	
		4	0.67	0.88	
		5	0.73	0.74	
		6	0.62	0.90	
		7	0.73	0.89	
Spanish	F722	1	0.65	0.83	878
		2	0.55	0.80	
		3	0.52	0.85	
		4	0.71	0.86	
		5	0.75	0.69	
		6	0.58	0.89	
		7	0.65	0.89	
French	F704	1	0.69	0.75	948
		2	0.64	0.87	
		3	0.64	0.75	
		4	0.74	0.64	
		5	0.41	0.57	
		6	0.65	0.61	
		7	0.71	0.69	
		8	0.77	0.76	
		9	0.56	0.58	
		10	0.40	0.72	
		A	0.63	0.82	
		B	0.65	0.80	
		C	0.63	0.86	
German	F714	1	0.66	0.67	422
		2	0.66	0.90	
		3	0.88	0.52	
		4	0.85	0.48	
		5	0.67	0.80	
		6	0.66	0.82	
		7	0.66	0.51	
		8	0.67	0.74	
		9	0.81	0.73	
		10	0.68	0.60	
		B	0.69	0.87	
		C	0.63	0.92	

Spanish	F724	1	0.61	0.64	604
		2	0.65	0.83	
		3	0.54	0.69	
		4	0.58	0.69	
		5	0.48	0.58	
		6	0.76	0.39	
		7	0.30	0.75	
		8	0.71	0.60	
		9	0.45	0.71	
		10	0.57	0.75	
		A	0.74	0.73	
		B	0.72	0.80	
		C	0.56	0.90	

To investigate the effectiveness of the assessments to measure candidates in relation to the grade boundaries, the partial credit model was fitted to the item-level data for these units. The very high tariff of items in section C (45 marks) are incompatible with this model and therefore marks for this section have been excluded from this part of the analysis.[22] Therefore, test information functions resulting from applying this model across sections A and B only are shown in Figure 4.5. Superimposed on these information functions are the distributions of the ability of candidates in latent variable space, along with the equivalent position of the grade boundaries. This shows that, for all three exams, there is some difference between the location of the grade boundaries and the information extracted by the items making up the assessment. This suggests that the targeting of the assessment is suboptimal (more pronounced for French and German). It is important to note that the omission of section C from this analysis is likely to have, if anything, slightly accentuated the difference between the location of the information function and the location of the grade boundaries. This is suspected to be the case as section C is one of the more demanding sections of the exam (see item facility indices in Table 4.2), with a high intended weighting. However, it is very unlikely that this section is sufficiently demanding to address this issue. While this difference is not extreme, given the relatively high facility indices observed across the written assessments, recommendations regarding targeting of the demand of items are made in section 10.

Figure 4.4: Unit-level raw mark distributions for the summer 2013 OCR A2 written exams for candidates certificating at A level in the same series

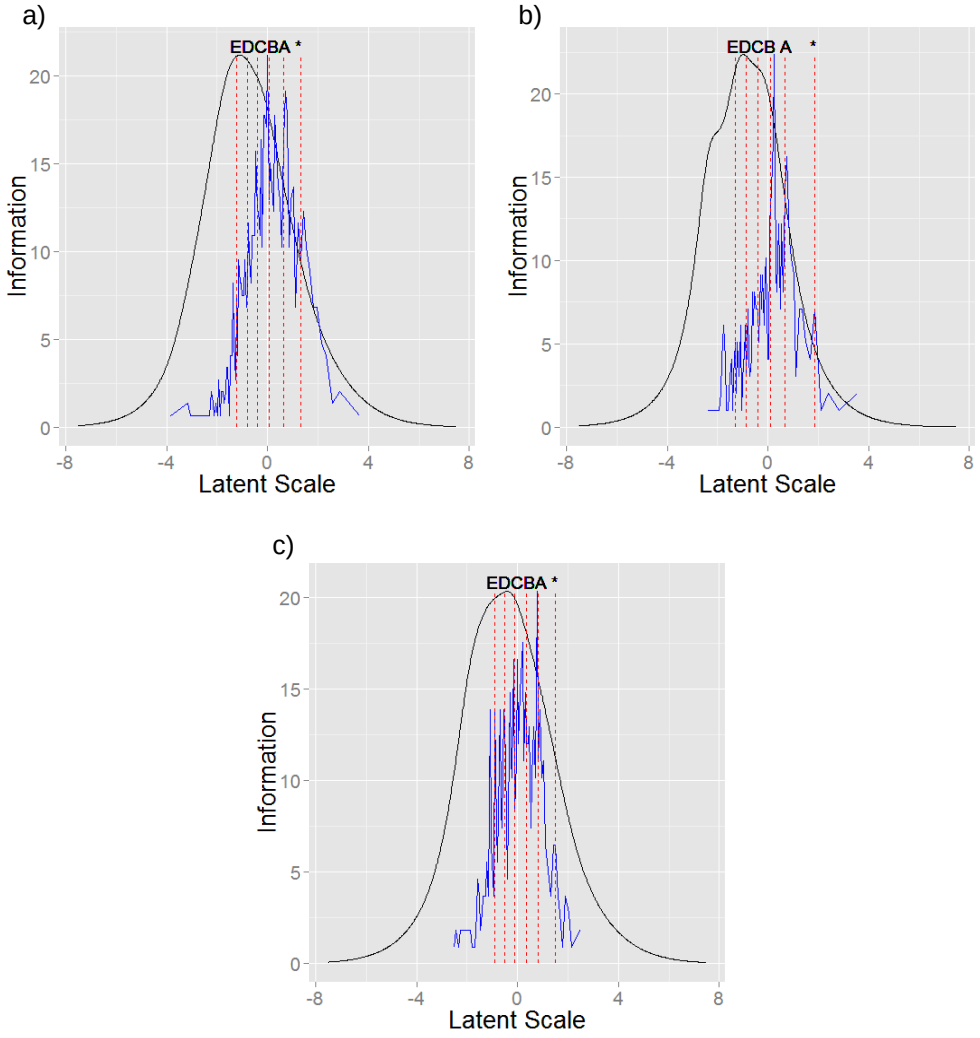


[22] Expected and empirical item-level category probability curves and item characteristic curves are provided in Appendices I, J and K for units F704, F714 and F724, respectively.

Table 4.3: Descriptive statistics for units F704, F714 and F724 from June 2013 for certificating A level candidates only

Unit	Mean		Standard deviation		Skewness
F704	90.0	(64.3%)	21.5	(15.4%)	-0.49
F714	94.3	(67.4%)	23.7	(16.9%)	-0.51
F724	80.1	(57.2%)	21.9	(15.6%)	-0.25

Figure 4.5: Test information functions (black) for a) F704, b) F714 and c) F724 from June 2013. Superimposed are the unit-level grade boundaries (dotted) and the distribution of candidate person parameters relative to these information functions (blue).



4.4 Exploration of item-level qualitative findings

The two items flagged by the subject expert reviewers were item 3e on unit F722 and item 7 on unit F724 (see Appendix B for full findings).

The facility index for item 3e on F722 (a nine-part question of objectively marked items) is lower than any across the AS written exams. It would be highly speculative, however, to suggest that this were solely due to the issues raised by the expert reviewers on a sub-part of the item. It is not possible to explore further the functioning of this sub-part due to the granularity of the data available.

The facility index for item 7 on unit F724 is the lowest on the question paper. However, again, the finding of the expert reviewers only affects a part of the overall item and therefore it would be speculative to suggest that this reduced facility index were due to the point raised.

While the specific points raised by the reviewers cannot be demonstrably linked to the functioning of the items, they are valid concerns and may be indicative of a wider issue with design of the items/mark schemes. It is recommended that the design of the mark scheme, specifically the basis on which elements of the translation are and are not considered acceptable, is reconsidered and the outcome clearly articulated. This would benefit future assessment development and would support greater transparency of the approach. Moreover, for items such as 7 on unit F724, having clarity over the principles underlying the mark scheme and ensuring that examiners have visibility and understanding of those principles as part of standardisation, will enable more consistent handling of responses not explicitly covered in the mark scheme.

4.5 Rank ordering of candidates

Above, the demand of exam items was explored through facility indices. While this provides information about the targeting of items, it does not provide any information regarding the validity of the rank order of candidates in the mark distribution. This issue is considered here for the OCR assessments through examination of the correlation of candidates' marks achieved on items assessing equivalent skills at AS and A2 level. The allocation of items to skill areas for the OCR assessments is provided in Table 4.3, with the intra-skill correlations provided in Table 4.4. These values are based on candidates who sat the AS written unit in summer 2012 and certificated at A level in summer 2013, as well as taking the A2 written assessment in that same series.

The scatter plots showing these relationships are provided in Figure 4.6.[23]

All listening items require candidates to provide short responses[24] with some scope for subjectivity in the marking. The quality of that marking could vary depending on the quality of the mark scheme and standardisation process. The lower correlation between marks for listening in Spanish compared to French and German is therefore worthy of further investigation beyond the current study.

The reading and reading/writing items are a mixture of objectively marked items, short-response items with some scope for subjectivity in the marking, and a transfer-of-meaning translation task. Given the nature of these items, the correlations for this skill appear satisfactory.

All written items require extended responses and are marked using a levels-of-response mark scheme. The AS items have two separate mark schemes for the content of the two sub-parts of item 7 (10 marks and 20 marks), with quality of language (accuracy) (10 marks) and quality of language (range) (10 marks) being awarded across the two written responses. The A2 marks are divided between relevance and points of view (10 marks), structure and analysis (15 marks), quality of language (accuracy) (10 marks) and quality of language (range) (10 marks). Given the scope for legitimate variation in marking within the levels-of-response mark schemes and the size of the correlations between the other skills, the correlations between writing tasks at AS and A2 appear to indicate a satisfactory level of marking quality.

[23] With the exception of German listening, the scatter plots show limited evidence of a ceiling effect. Calculating modified correlation coefficients using the method outlined in section 3.6 produced modified coefficients that are negligibly different from those presented in Table 4.4, with the exception of German listening, which produced a modified coefficient of 0.55 (n = 83).

[24] Maximum item sub-part tariff is four marks with a typical tariff of one or two marks.

Table 4.3: Designation of items to skills for the OCR written assessments

Item number	June 2012	June 2013	
	F702 / F712 / F722	F704 / F724	F714
1	L	L	L
2	L	L	L[25]
QoL_L[26]	-	L	-
3	L	R	R
4	RW	RW	RW
5	R	RW	RW
6	R	RW	RW
7	WC	RW	RW
8	-	RW	RW
9	-	RW	RW
10	-	RW	RW
QoL_RW	-	RW	RW
Section C[27]	-	WO	WO

[25] In contrast to F704 and F724, the quality of language marks for section A of F714 were supplied aggregated with those for item 2.

[26] While this item is designated as listening in the supplied data, the marks are awarded for the quality of the written response and are therefore excluded from analysis of the relationship between listening marks.

[27] In the supplied data, section C is designated as WC (written compulsory). There are, however, a number of optional questions in this section and it is therefore assigned here as WO.

Table 4.4: Correlation between candidates' marks in equivalent skills in AS and A2 assessments for candidates certificating with OCR at A level in summer 2013

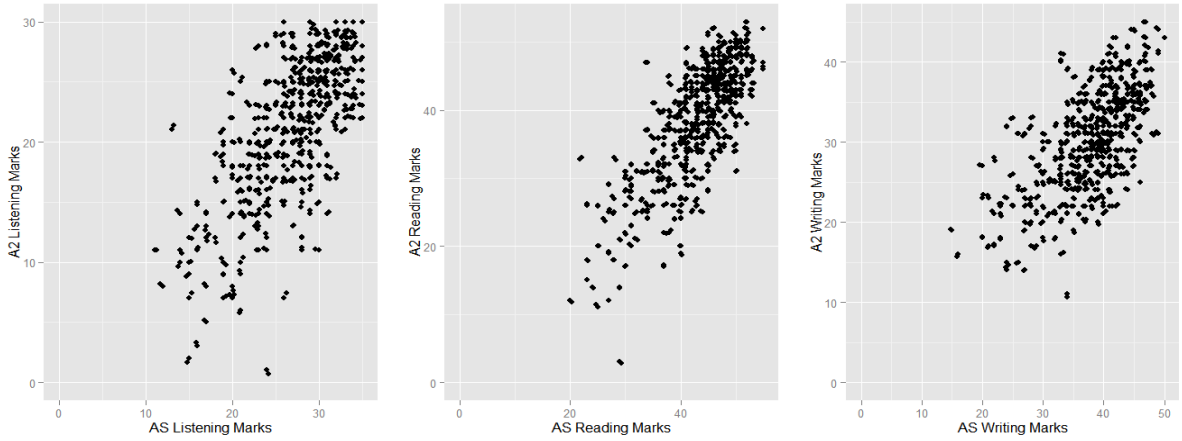
	Skill	Items used for comparison		Item totals	SD – % of item(s) max	A2:AS correlation (zeros removed[28])	No. of candidates
French	Listening	AS	Q1 Q2 Q3	35	14.2	0.72	427
		A2	Q1 Q2	30	23.7		
	Reading & R/W	AS	Q4 Q5 Q6	55	12.3	0.78	428
		A2	Q3–Q10	55	20.9		
	Writing	AS	Q7	50	12.7	0.65	428
		A2	Sec C	45	18.6		
German	Listening	AS	Q1 Q2 Q3	35	12.0	0.73	164
		A2	Q1 Q2[29]	35	24.0		
	Reading & R/W	AS	Q4 Q5 Q6	55	12.6	0.80	164
		A2	Q3–Q10	55	21.6		
	Writing	AS	Q7	50	13.7	0.63	164
		A2	Sec C	45	23.2		
Spanish	Listening	AS	Q1 Q2 Q3	35	16.1	0.66	245
		A2	Q1 Q2	30	21.3		
	Reading & R/W	AS	Q4 Q5 Q6	55	11.7	0.68	245
		A2	Q3–Q10	55	22.3		
	Writing	AS	Q7	50	17.2	0.75	245
		A2	Sec C	45	22.5		

[28] The correlation coefficients are evaluated with candidates removed who scored a total of zero marks for either all AS or all A2 items in the skill to exclude candidates who likely entered no response.

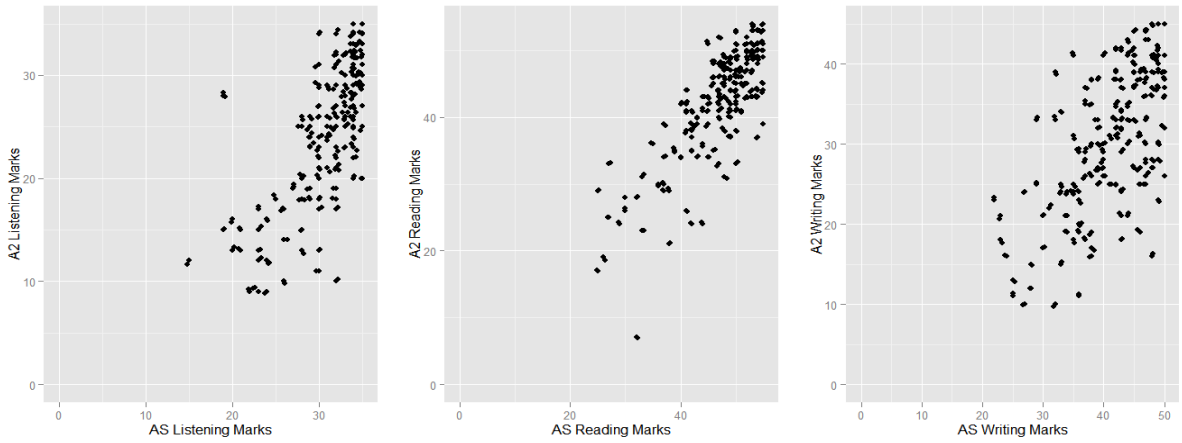
[29] Note that the German A2 listening marks contain quality of language marks in item 2.

Figure 4.7: Scatter plots[30] showing OCR candidates' marks for listening, reading and writing skills in French, German and Spanish at AS and A2 level for candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

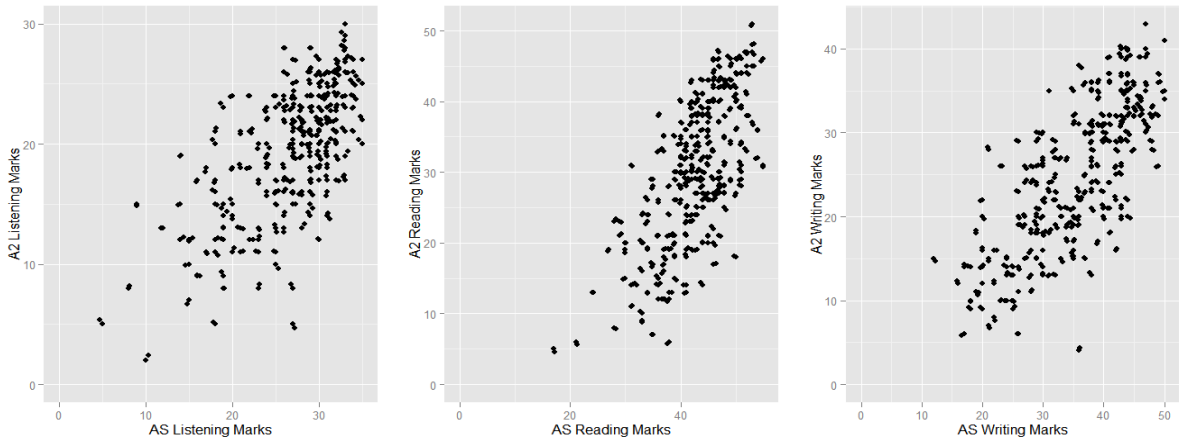
a) French



b) German



c) Spanish



[30] The data points have been jittered to avoid over-plotting to show the density of points.

4.6 Weighting of skills

As discussed in section 3.7, the differential functioning of items between skills may impact on the extent to which those skills contribute to the overall ranking of candidates. The achieved and intended weightings of the skills on the OCR A2 written assessments are presented in Table 4.5. While there are some relatively large differences between the achieved and intended weights of the skills (listening and writing) in the French assessment (F704), the differences are, in general, not at a concerning level. These values should be monitored over time in order to ensure that any differences between intended and achieved weight are neither systematic nor increasing over time.

Table 4.5: Achieved vs intended weight for OCR A2 written assessments

Unit	Skill designation	Intended weight %	Achieved weight %	Difference (% pnts)
F704	L	25.0	30.6	+5.6
	R	5.0	6.5	+1.5
	RW	37.9	36.4	-1.5
	W	32.1	26.5	-5.6
F714	L	25.0	27.3	+2.3
	R	5.7	2.9	-2.8
	RW	37.1	36.4	-0.7
	W	32.1	33.4	+1.3
F724	L	25.0	24.1	-0.9
	R	2.9	1.6	-1.3
	RW	40.0	42.3	+2.3
	W	32.1	32.0	-0.1

4.7 Speaking assessments

In Figure 4.8 are the unit-level mark distributions for the OCR A2 speaking tasks (units F703, F713 and F723 for French, German and Spanish, respectively). Descriptive statistics summarising these distributions are provided in Table 4.6. These distributions have high mean marks awarded to candidates, with around one-third of the lower end of the mark distribution being largely unused. Despite these high mean marks, the only unit that appears to have a truncated distribution, and therefore likely to result in an impaired discrimination between the most able candidates, is the German unit F713. However, the functioning of all of these assessments would benefit from the expectations of candidates (through modification of the mark scheme or through the standardisation process) being raised to spread candidates more effectively across the mark distribution.

Findings 7b and 8a from the expert review (see Appendix B) suggested that the expectations of the top mark band were very high. However, given the shapes of

these mark distributions, it appears that candidates are either meeting these expectations or the expectations are being modified to reflect the context in which the performance is delivered.

Similar to the analysis performed in section 4.5, Figure 4.9 and Table 4.7 summarise the relationship between candidates' spoken language marks at AS in June 2012 (F701, F711 and F721) and A2 in June 2013 (F703, F713 and F723). Setting these correlations against those reported in Table 4.4, given the challenges of marking assessments of this kind, these correlations appear satisfactory.

Table 4.6: Descriptive statistics for units F703, F713 and F723 from June 2013 for certificating A level candidates only

Unit	Mean		Standard deviation		Skewness
F703	41.5	(69.2%)	7.7	(12.8%)	0.05
F713	47.0	(78.3%)	9.9	(16.5%)	-0.77
F723	45.5	(75.8%)	8.1	(13.5%)	-2.04[31]

[31] Note that this high value of negative skew arises from the relatively high proportion of candidates certificating with zero marks for F723. Excluding these candidates, this value of skewness is -0.31.

Figure 4.8: Unit-level raw mark distributions for F703, F713 and F723 for certificating A level candidates in summer 2013 with grade boundaries superimposed

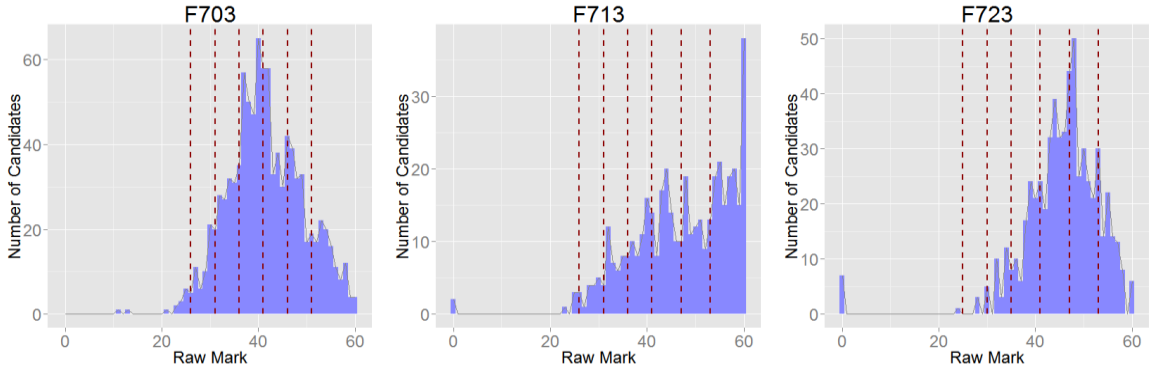
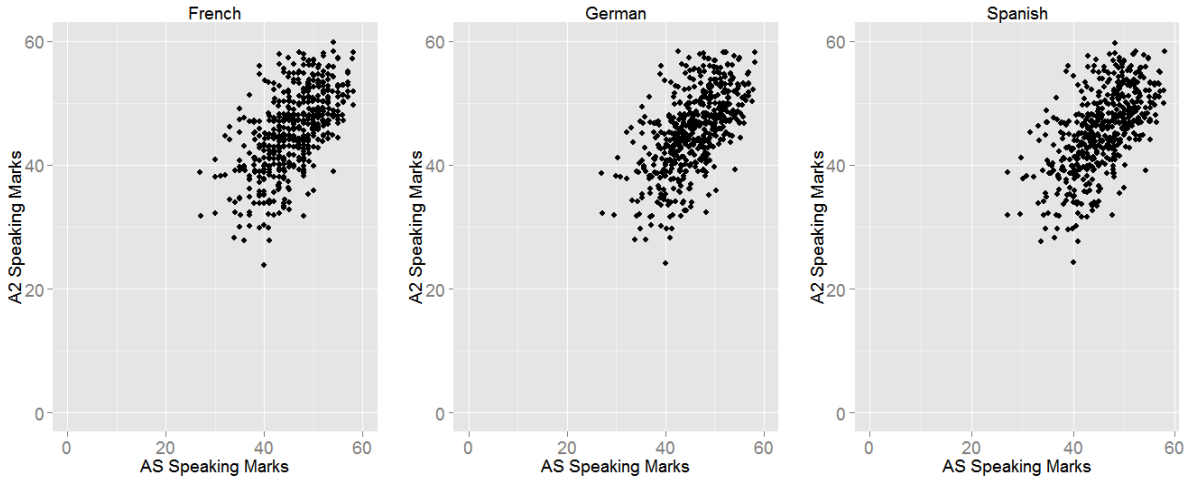


Figure 4.9: Scatter plots[32] showing candidates' marks for speaking in French, German and Spanish at AS and A2 level for OCR candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013



[32] The data points have been jittered to avoid over-plotting to show the density of points.

Table 4.7: Correlation between candidates' marks achieved in speaking at AS level (unit 1) and A2 level (unit 3)

	AS:A2 correlation	No. of candidates in correlation
French	0.62	874
German	0.72	349
Spanish	0.60	529

5 Pearson

5.1 Assessment structure

The assessment structure for the A level MFL specifications offered by Pearson is summarised in Table 5.1.

Table 5.1: Pearson A level MFL assessment framework

Level	Unit code	Mode of assessment	Intended weight within A level	Assessment objectives	Max raw mark
AS	6FR01 6GN01 6SP01	Speaking tasks	15%	AO1 = 10% AO3 = 5%	50
	6FR02 6GN02 6SP02	Written examination	35%	AO1 = 10% AO2 = 17.5% AO3 = 7.5%	70
A2	6FR03 6GN03 6SP03	Speaking tasks	17.5%	AO1 = 12.5% AO2 = 2.5% AO3 = 2.5%	50
	6FR04 6GN04 6SP04	Written examination	32.5%	AO2 = 22.5% AO3 = 10%	100

5.2 Subject expert scrutiny

A consolidated version of the findings from the qualitative review of Pearson's assessment materials is provided in Appendix C. The points suitable for quantitative exploration using the available data are:

> 7a. Some of the phrases used in band descriptors for the essay questions on units 6SP02 and 6SP04 appear to set very high expectations. For example, 6SP02, question 8, the top band for the content and response grid has "Task fully grasped, answer wholly relevant...." The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for a 17- or 18-year-old candidate should look like at the very highest level.
>
> 8a. For units 2 and 4, the very high expectations generated by phrasing used in some of the top mark bands for the essay questions may advantage native speakers if the understanding of what a top performance

> from a non-native-speaking 17- or 18-year-old candidate looks like is not consistent.

Finding 7a will be addressed in section 5.3. Finding 8a impacts on the analyses presented in section 5.4.

5.3 Item functioning

The question structures within the A2 units of the Pearson specifications are considerably different to those offered by the other exam boards. Candidates complete only three items (two of which are from a choice of options) in the written A2 assessments:

> Q01 An item which requires candidates to translate a passage from English into the target language (10 marks – although an initial allocation of 30 marks is made with a 0.33 scaling factor applied);
>
> Q02 An item with a number of options. Candidates produce a creative or discursive essay (45 marks; 15 marks for Understanding and Response, 15 marks for Organisation and Development, 10 marks for Range and Application of Language, 5 marks for the Accuracy of Language);
>
> Q03 An extended written research based essay chosen from a number of options (45 marks; 30 marks for Reading research and understanding, 9 marks for Organisation and development, 6 marks for Quality of language).

The downward scaling applied to the marks allocated to item 1 is unsatisfactory as it inevitably results in the loss of information. The rationale for initially marking at a high level of resolution (30 marks) but then scaling to a much lower resolution (10 marks) should be reconsidered.

Despite the subdivision of marks between different areas of the mark scheme for items 2 and 3, the data available are aggregated to the level of the overall item. This means that examination of the functioning of the different elements of these marks and their interrelationship is not possible here. No indication is provided in the mark scheme that there is an imposed interdependency on the marks awarded for the various elements of these items. The item-level mark distributions for the summer 2013 A2 written exams for certificating A level candidates are shown in Figures 5.1, 5.2 and 5.3, with the item-level descriptive statistics for both AS and A2 written exams provided in Table 5.2.

Figure 5.1: Item-level mark distributions for 6FR04 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

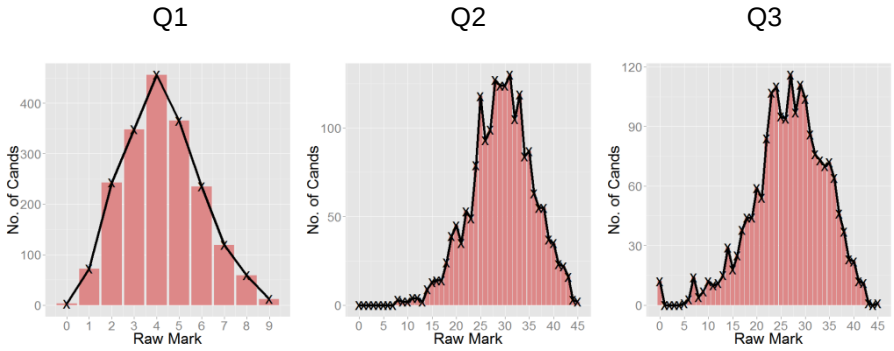


Figure 5.2: Item-level mark distributions for 6GN04 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

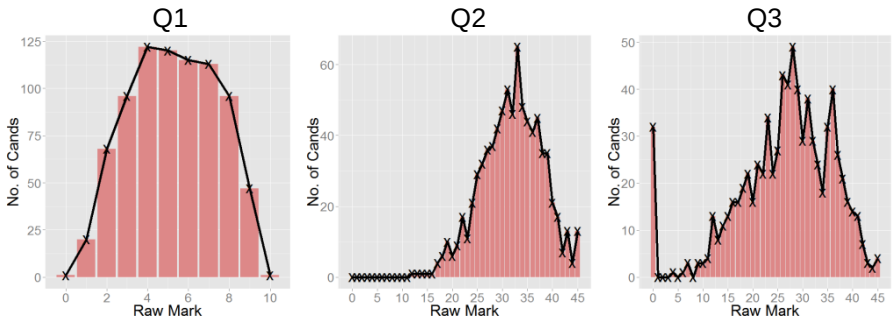


Figure 5.3: Item-level mark distributions for 6SP04 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

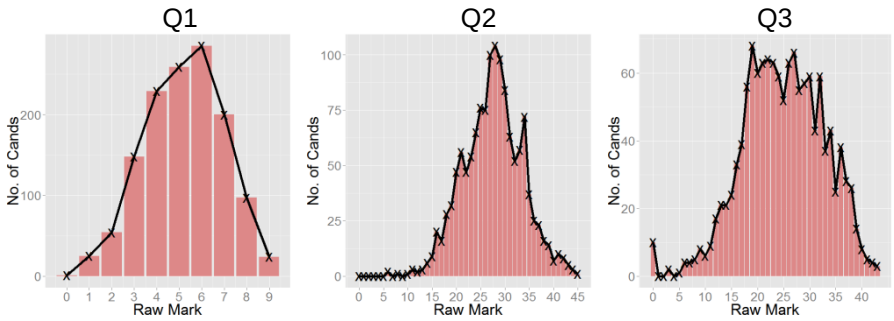


Table 5.2: Item facility indices for 6FR04, 6GN04 and 6SP04 for certificating candidates in June 2013

	Unit	Item reference	Facility index	Discrimination index	No. of candidates
French	6FR02	1	0.73	0.64	2,520
		2	0.76	0.55	
		3	0.72	0.67	
		4	0.30	0.77	
		5	0.90	0.53	
		6	0.35	0.55	
		7	0.43	0.80	
		8	0.66	0.86	
	6FR04	1	0.47	0.69	1,912
		2	0.65	0.82	
		3	0.59	0.85	
German	6GN02	1	0.54	0.56	1,391
		2	0.80	0.60	
		3	0.63	0.58	
		4	0.55	0.75	
		5	0.87	0.56	
		6	0.69	0.50	
		7	0.51	0.69	
		8	0.69	0.89	
	6GN04	1	0.53	0.66	799
		2	0.71	0.73	
		3	0.59	0.86	
Spanish	6SP02	1	0.80	0.63	2,120
		2	0.80	0.64	
		3	0.73	0.69	
		4	0.64	0.78	
		5	0.79	0.63	
		6	0.51	0.75	
		7	0.64	0.76	
		8	0.68	0.87	
	6SP04	1	0.58	0.55	1,322
		2	0.61	0.71	
		3	0.58	0.81	

In all three languages, the translation task (question 1) discriminates well, with all regions of the mark scale being used. The same is true for the research writing task (question 3). Question 2 does not, however, use the full extent of the mark range, with the lower 12 marks being largely unused. Due to the granularity of the data currently available, it is not possible to determine whether a particular element of the allocated mark is leading to this region of the mark scale being unused. Given that discrimination on these assessments is largely achieved by outcome rather than task, pending deeper exploration of the data, it appears that a greater spread of marks for question 2 could be achieved post hoc through the examiner standardisation process.

Despite this effect, initial consideration of these item distributions appears to suggest that they are functioning relatively well. Aggregating candidates' marks to unit level leads to the unit-level mark distributions as shown in Figure 5.4. The combined effect of these item distributions is a reasonably small area of unused mark scale at the lower end of the distribution arising from the underuse of this region of the mark scale in question 2. The descriptive statistics summarising these distributions are provided in Table 5.3. Given the high tariff of these items and the relatively good spread of marks across the mark scale, analysis of the test information using the partial credit model is neither necessary nor appropriate.

Figure 5.4: Unit-level raw mark distributions for the summer 2013 Pearson A2 written exams for candidates certificating at A level in the same series

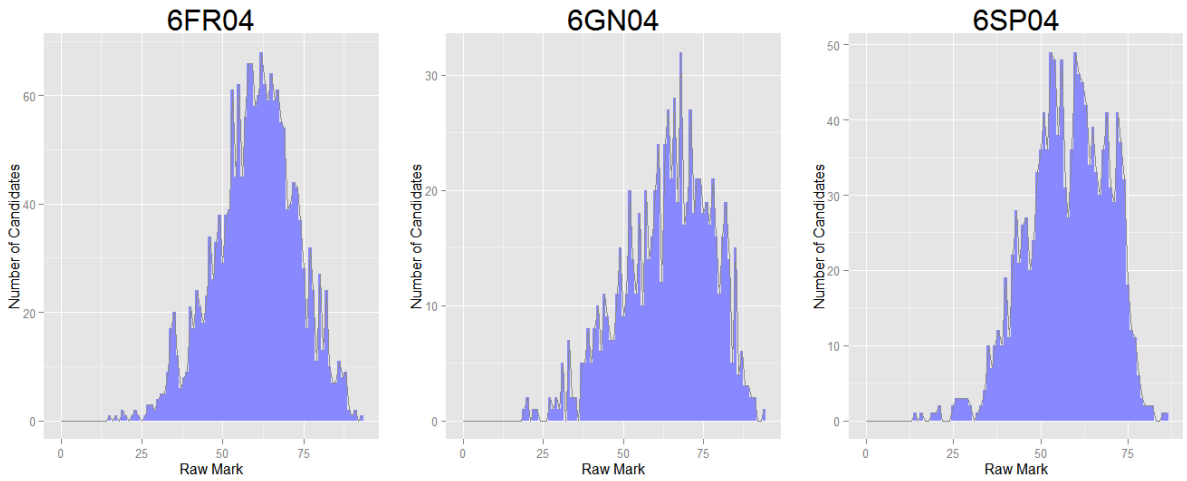


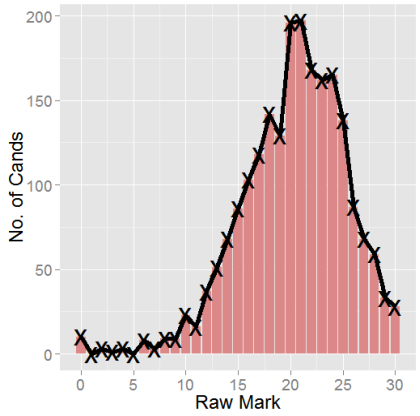
Table 5.3: Descriptive statistics for units 6FR04, 6GN04 and 6SP04 from June 2013 for certificating A level candidates only

Unit	Mean		Standard deviation		Skewness
6FR04	57.7	(57.7%)	11.5	(11.5%)	-0.40
6GN04	60.3	(60.3%)	12.6	(12.6%)	-0.27
6SP04	63.7	(63.7%)	13.9	(13.9%)	-0.49

5.4 Exploration of expert reviewer findings

Finding 7a raised concerns over the expectations of candidates if they are to achieve the highest marks on questions such as item 8 of unit 6SP02. In Figure 5.5 is the item-level mark distribution for this item from the summer 2013 exam series. Based on this distribution and those observed in Figures 5.1 to 5.3, while the statements appearing in the mark scheme set very high expectations, either a number of candidates are successfully meeting these expectations or the context in which candidates are encountering the items is appropriately reflected in the examiner standardisation process resulting in some candidates being awarded top marks.

Figure 5.5: Item-level raw mark distribution for item 8 of unit 6SP02 from the summer 2013 exam series



5.5 Rank ordering of candidates

The mark scheme construction for questions 2 and 3 in units 6GN04 and 6SP04 consists of a brief outline of the anticipated response accompanied by a levels-of-response mark scheme. For question 2, separate levels of response-marking grids are available for understanding and response (15 marks), organisation and development (15 marks), range and application of language (10 marks) and accuracy of the target language (5 marks). For question 3, banded mark schemes for reading research and understanding (30 marks), organisation and development (9 marks) and quality of language (6 marks) are used. Due to the absence of the levels-of-

response marking grids in the mark scheme for unit 6FR04, it is unclear whether or not this approach was used to support the marking of these questions.

The intra-skill correlation coefficients between AS and A2 assessments are provided in Table 5.4 for candidates sitting the AS written exam in summer 2012, the A2 written exam in summer 2013 and also certificating at A level in summer 2013. Only reading and writing skills are reported here as listening is not assessed in the written A2 assessments in the Pearson specifications. The corresponding scatter plots are provided in Figure 5.6.

The AS reading items are either multiple choice or very short response, so a high level of marking reliability is likely. The AS writing items are extended responses between 200 and 220 words marked using a two-part levels-of-response mark scheme, with a maximum of 15 marks awarded for content and response and a maximum of 15 marks for quality of language.

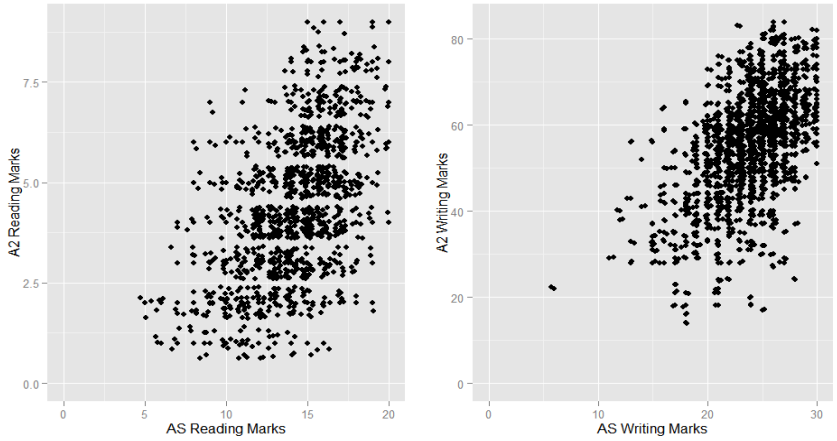
Table 5.4: Correlation between candidates' marks in equivalent skills in AS and A2 assessments for candidates certificating with Pearson at A level in summer 2013

	Skill	Items used for comparison		Item totals	SD – % of item(s) max	A2:AS correlation (zeros removed[33])	No. of candidates
French	Reading	AS	Q05 Q06 Q07	20	13.2	0.55	1,133
		A2	Q01	10	22.5		
	Writing	AS	Q08	30	11.3	0.54	1,134
		A2	Q02 Q03	90	22.3		
German	Reading	AS	Q05 Q06 Q07	20	16.4	0.56	303
		A2	Q01	10	25.4		
	Writing	AS	Q08	30	11.7	0.50	303
		A2	Q02 Q03	90	26.4		
Spanish	Reading	AS	Q05 Q06 Q07	20	15.9	0.49	585
		A2	Q01	10	25.2		
	Writing	AS	Q08	30	11.4	0.32	578
		A2	Q02 Q03	90	23.7		

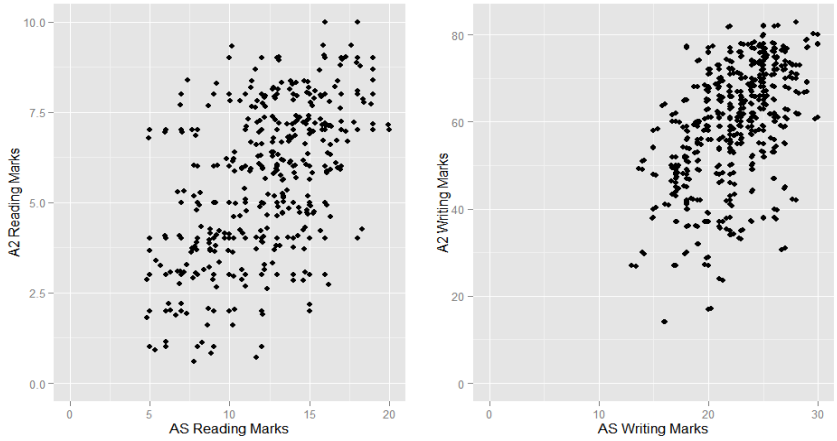
[33] The correlation coefficients are evaluated with candidates removed who scored a total of zero marks for either all AS or all A2 items in the skill to exclude candidates who likely entered no response.

Figure 5.6: Scatter plots[34] showing Pearson candidates' marks for reading and writing skills in French, German and Spanish at AS and A2 level for candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

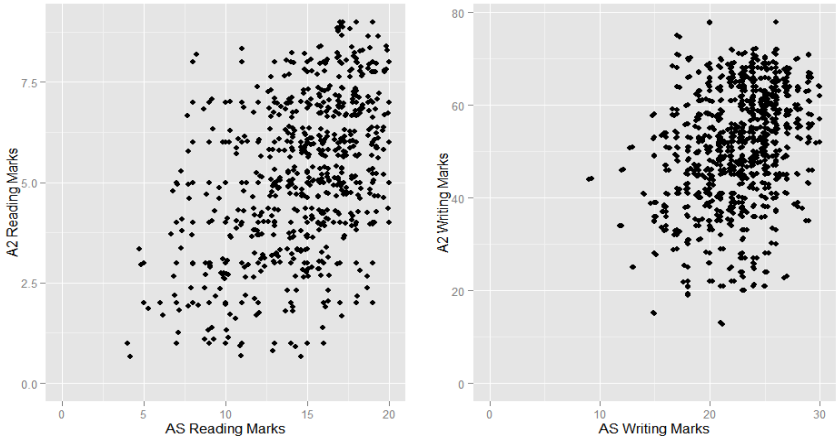
a) French



b) German



c) Spanish



[34] The data points have been jittered to avoid over-plotting to show the density of points.

Other than the AS assessment of writing in French, there is little evidence of a strong ceiling effect in these distributions (see Figure 5.6).[35] In light of this, and given the objective nature of the marking of the AS items, the correlations between marks at AS and A2 are surprisingly low. Similar to the points raised in section 3.7, the source of these relatively low correlations may be:

1. greater variability in the progression of candidates in their reading and writing skills between AS and A2 than in reading and listening;

2. unexpectedly low marking reliability;

3. a relatively low assessment reliability due to the length of the assessment;

4. the combination of marks across optional routes through the question papers that may be indicative of variations in demand of those optional routes.

The data currently available are not sufficient to enable potential sources 2 to 4 to be decoupled from one another. Therefore, a further investigation is required to determine the extent to which these potential sources of reduced correlation (and therefore potentially reduced validity of the mark distribution) contribute. This evidence in itself does not prove that marking reliability is poor or that there is a lack of comparability between optional routes. However, it is sufficiently concerning to warrant further investigation using more in-depth operationally available data.

5.6 Weighting of skills

Shown in Table 5.5 are the intended and achieved weights of the skills assessed in the Pearson A2 written units. This shows that there is a very close match between the intended and achieved weight of the skills. It is important to note, however, that this in no way detracts from the need to perform the additional analyses outlined in section 5.4 since these metrics only indicate a consistency of functioning of items across skills rather than anything directly about the quality of those elements of the assessment.

[35] Due to the absence of a strong ceiling effect in these data, the modified correlation coefficients (as described in section 3.7) are not reported here for all languages/skills. However, the relationship with the strongest apparent ceiling effect is French writing, giving rise to a modified correlation coefficient of 0.51 (n = 982).

Table 5.5: Achieved vs intended weight for Pearson A2 written assessments

Unit	Skill designation	Intended weight %	Achieved weight %	Difference (% pnts)
6FR04	R	10.0	9.3	-0.7
	W	90.0	90.7	+0.7
6GN04	R	10.0	10.1	+0.1
	W	90.0	89.9	-0.1
6SP04	R	10.0	8.3	-1.7
	W	90.0	91.7	+1.7

5.7 Speaking assessments

Figure 5.7 shows the raw mark distributions for the Pearson A2 speaking assessments, with the corresponding descriptive statistics provided in Table 5.6. Given the high prevalence of candidates certificating with zero marks on this unit from this series, recalculated summary statistics with these candidates excluded are presented in Table 5.7.

Figure 5.7: Unit-level raw mark distributions for 6FR03, 6GN03 and 6SP03 for certificating A level candidates in summer 2013 with grade boundaries superimposed

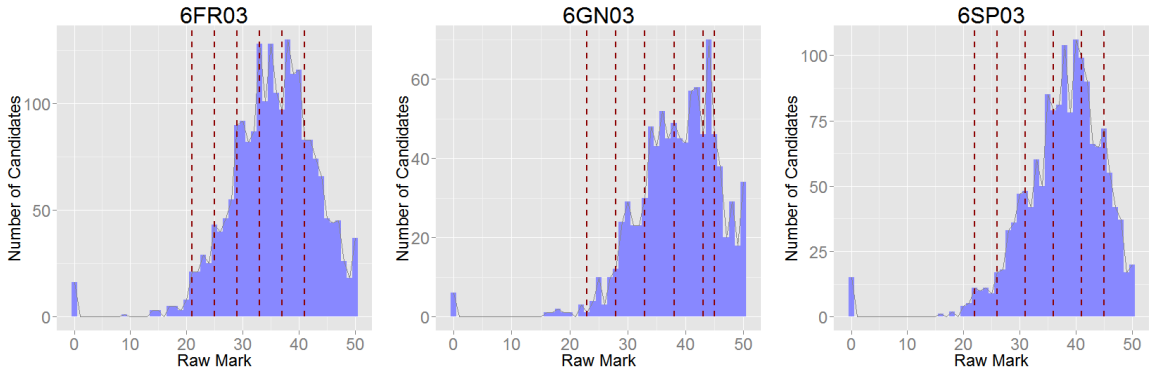


Table 5.6: Descriptive statistics for units 6FR03, 6GN03 and 6SP03 from June 2013 for certificating A level candidates only

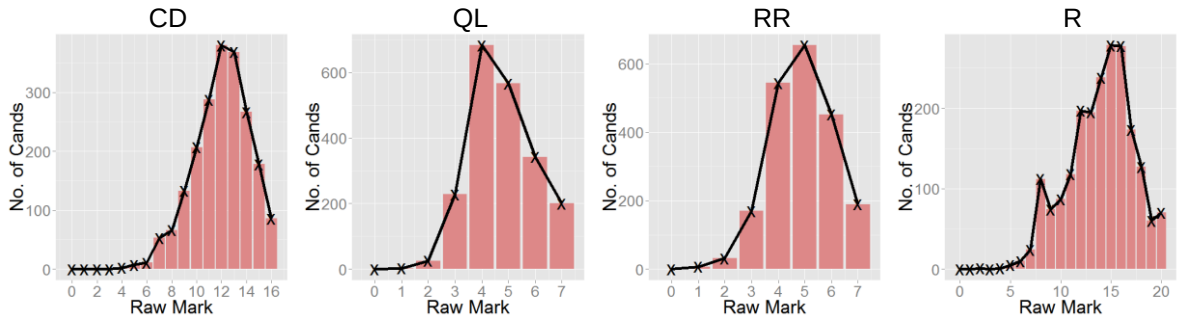
Unit	Mean		Standard deviation		Skewness
6FR03	35.4	(70.1%)	7.5	(15.0%)	-0.85
6GN03	38.7	(77.4%)	7.1	(14.2%)	-1.28
6SP03	37.6	(75.2%)	7.3	(14.6%)	-1.50

Table 5.7: Descriptive statistics for units 6FR03, 6GN03 and 6SP03 from June 2013 for certificating A level candidates only (zero marks excluded)

Unit	Mean		Standard deviation		Skewness
6FR03	35.7	(71.4%)	6.9	(13.8%)	-0.20
6GN03	39.0	(78.0%)	6.5	(13.0%)	-0.42
6SP03	38.0	(76.0%)	6.3	(12.6%)	-0.44

These distributions show that candidates are scoring very highly on these assessments, with the lowest 20 marks being largely unused. The availability of sub-task-level data for the Pearson speaking task means that the composition of these unit-level mark distributions can be explored and any items likely contributing to this effect identified. In Figures 5.7, 5.8 and 5.9 are the item-level mark distributions for units 6FR03, 6GN03 and 6SP03, respectively. Candidates are marked based on their comprehension and development (16 marks), quality of language (7 marks), reading and research (7 marks) and quality of response (20 marks) (characterised by spontaneity, handling of abstract concepts and the range of lexis and structures used).

Figure 5.7: Item-level mark distributions for 6FR03 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series[36]



[36] CD = comprehension and development, QL = quality of language, RR = reading and research, R = quality of response (for Figures 5.7 to 5.9).

Figure 5.8: Item-level mark distributions for 6GN03 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series

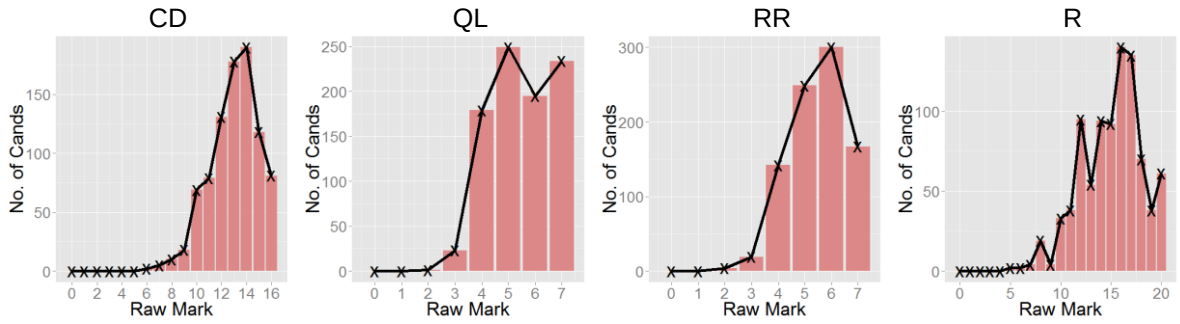
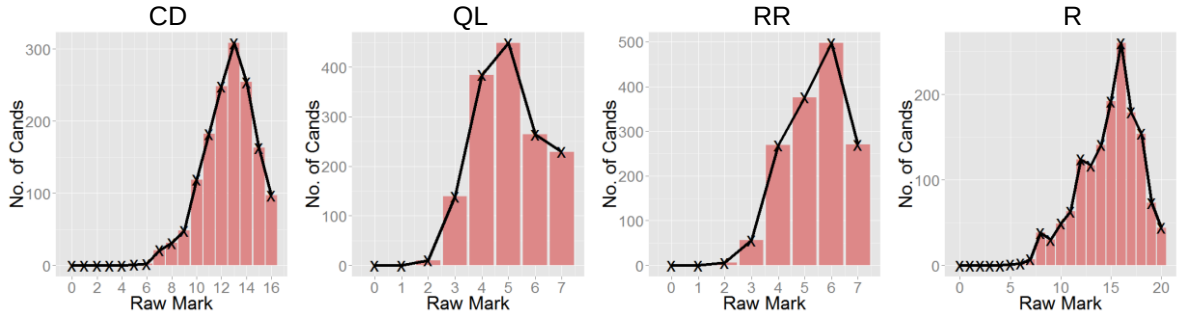


Figure 5.9: Item-level mark distributions for 6SP03 for candidates sitting the unit and certificating candidates at A level in the summer 2013 exam series



All of these mark distributions have an unused region at the low end of the mark distribution and therefore no particular item appears to be contributing to the compression of marks at unit level. This suggests that increasing the expectations of the quality of candidates' responses in all areas would aid the valid discrimination between candidates of different abilities.

Figure 5.10 shows the relationship between candidates' marks achieved on the AS and A2 speaking assessment for A level candidates who certificated in summer 2013 and entered unit 1 (AS speaking) in June 2012 and unit 3 (A2 speaking) in June 2013. The correlations between these marks are provided in Table 5.8. As with the intra-skill correlations for the skills assessed through the written exams, these correlations are relatively low. While the challenges of standardising and monitoring a large number of markers to assess candidates using a mark scheme that, necessarily, involves a level of subjectivity are recognised approaches to address, the lack of discrimination between candidates should be considered.

Figure 5.10: Scatter plots[37] showing candidates' marks for speaking in French, German and Spanish at AS and A2 level for Pearson candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

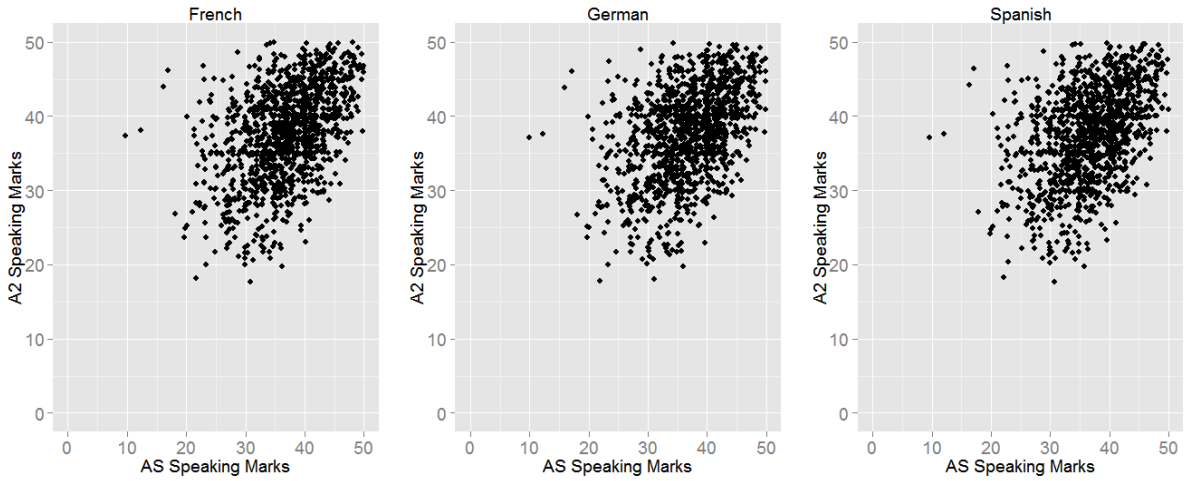


Table 5.8: Correlation between candidates' marks achieved in speaking at AS level (unit 1) and A2 level (unit 3)

	AS:A2 correlation	No. of candidates in correlation
French	0.51	1,844
German	0.49	706
Spanish	0.46	1,242

[37] The data points have been jittered to avoid over-plotting to show the density of points.

6 WJEC

6.1 Assessment structure

The assessment structure for the current A level MFL specifications offered by WJEC is summarised in Table 6.1.

Table 6.1: WJEC A level MFL assessment framework

Level	Unit code	Mode of assessment	Intended weight within A level	Assessment objectives	Max raw mark
AS	FN1 GN1 SN1	Speaking task (either assessed by a visiting examiner or centre assessed)	20%	AO1 = 16.7% AO3 = 3.3%	60
	FN2 GN2 SN2	Written examination	30%	AO1 = 2.4% AO2 = 18.4% AO3 = 9.2%	98
A2	SN3 GN3 SN3	External assessed speaking task	20%	AO1 = 10.0% AO2 = 6.7% AO3 = 3.3%	60
	FN4 GN4 SN4	Written examination	30%	AO1 = 4.9% AO2 = 15.9% AO3 = 9.2%	98

6.2 Subject expert scrutiny

Provided in Appendix D are the detailed findings of the review of WJEC assessment materials performed by the subject experts. Key findings that will be examined further here using the available data are:

> 2b. The mark schemes for several tasks do not include details of what alternative answers are acceptable and what will be rejected:
>
> SN2, reading tasks 3b and 5;
>
> SN4, listening tasks 1a and 1b, and reading tasks 2a and 2c.
>
> SN4 reading tasks 2a and 2c in particular require a high level of manipulation, inference, deduction and personal opinion, so markers need to be clear what responses are acceptable and what should be rejected.
>
> 7a. For SN4, some of the words/phrases used in band descriptors appear to set very high expectations. For example, for range and idiom, the top band describes “Assured sense of register. Uses language imaginatively to achieve desired effect. Evidence of style, nuance….” The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for an 18-year-old candidate should look like at the very highest level.

Consideration of the functioning of the items cited in finding 2b and the issue highlighted in finding 7a are considered in section 6.4.

6.3 Item functioning

Provided in Table 6.2 are the item summary statistics for all WJEC written exam units in summer 2013. The item-level raw mark distributions are provided in Figures 6.1, 6.2 and 6.3 for French, German and Spanish, respectively.

Table 6.2: Item-level summary statistics for WJEC AS and A2 units in summer 2013

	Unit	Item reference	Facility index	Discrimination index	No. of candidates
French	FN2	1	0.59	0.69	2,873
		2	0.58	0.74	
		3a	0.83	0.34	
		3b	0.34	0.65	
		3c	0.45	0.73	
		4a	0.52	0.64	
		4b	0.82	0.59	
		4c	0.89	0.52	
		5	0.67	0.78	
		6	0.68	0.84	
	FN4	1	0.63	0.60	1,951
		2a	0.41	0.74	
		2b	0.56	0.74	
		3	0.47	0.86	
		4	0.69	0.90	
German	GN2	1	0.52	0.71	1,241
		2	0.37	0.65	
		3.1	0.60	0.86	
		3.2	0.80	0.62	
		3.3	0.56	0.77	
		4	0.63	0.83	
		5.1	0.68	0.77	
		5.2	0.57	0.74	
		6	0.67	0.90	
	GN4	1	0.78	0.59	821
		2.1	0.81	0.72	
		2.2	0.90	0.39	
		2.3	0.88	0.53	
		2.4	0.89	0.41	
		3	0.63	0.91	
		4	0.63	0.94	
Spanish	SN2	1	0.68	0.73	1,893
		2	0.68	0.32	
		3a	0.78	0.42	
		3b	0.29	0.68	
		3c	0.55	0.86	
		4	0.64	0.73	
		5	0.37	0.84	

		6	0.62	0.91	
	SN4	1a	0.48	0.62	1,268
		1b	0.54	0.41	
		2a	0.50	0.66	
		2b	0.76	0.50	
		2c	0.38	0.78	
		3	0.52	0.90	
		4	0.66	0.93	

Figure 6.1: Item-level mark distributions for FN4 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

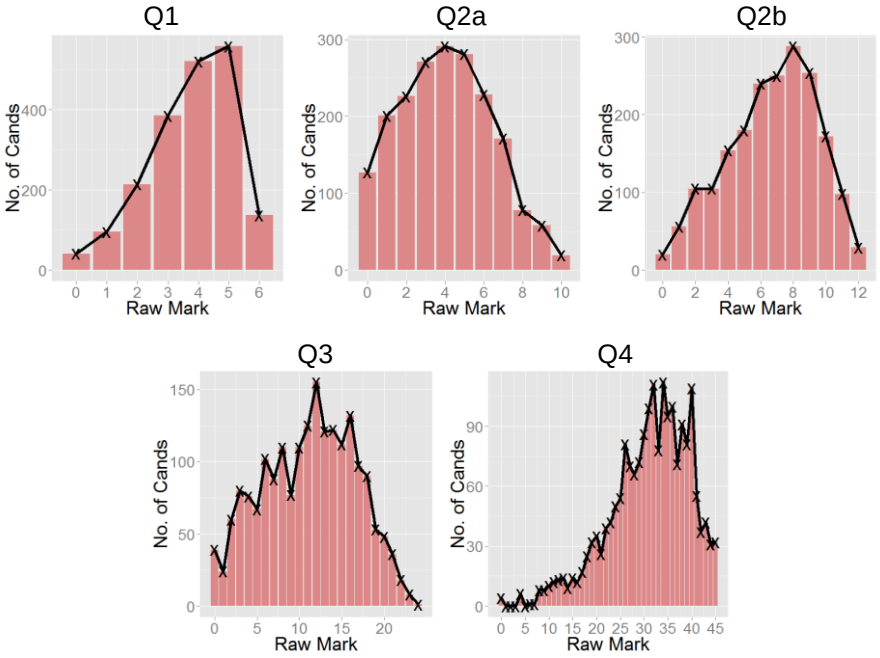


Figure 6.2: Item-level mark distributions for GN4 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

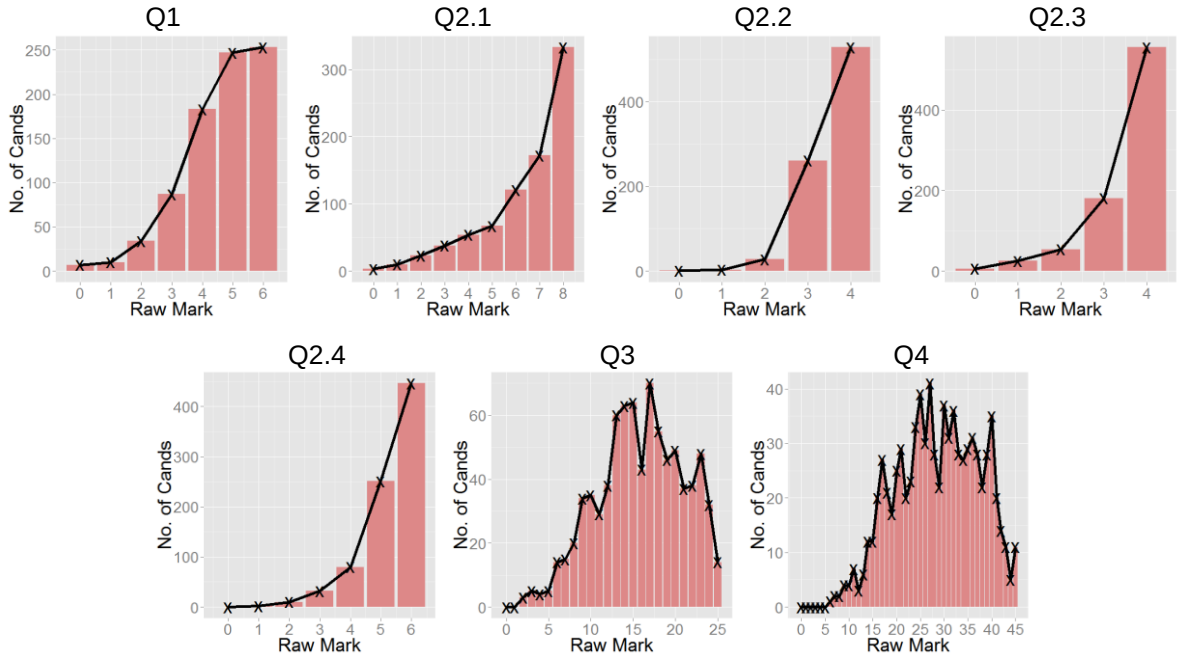
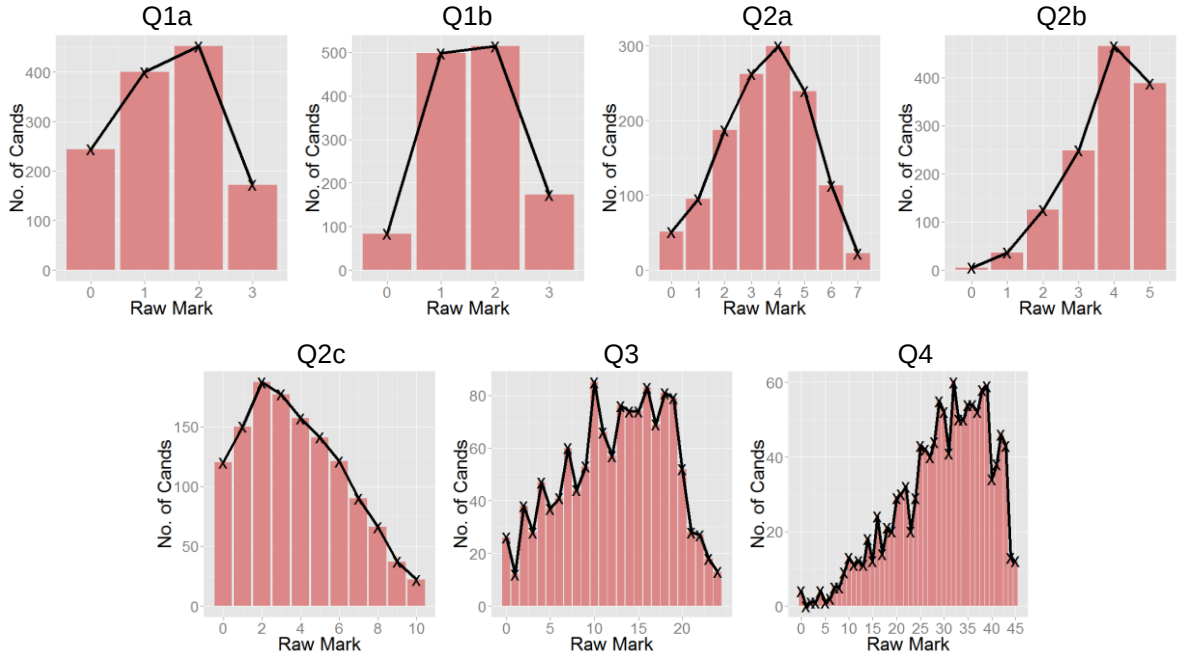


Figure 6.3: Item-level mark distributions for SN4 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series



For French and Spanish, the items spread candidates relatively well across the mark scale for the A2 units. A range of item facilities are observed across both AS and A2 written exams. This is not the case for the German assessments, however. With the exception of the writing tasks (questions 3 and 4), the facility indices for all items on the German A2 written exam are high, suggesting that these items are not as demanding for candidates sitting this assessment. The potential for the ability of candidates sitting German to be higher than the other languages may, however, be a cause of this effect as is explored further below.

The consequences of these item-level distributions at unit level can be seen in Figure 6.4. These unit-level mark distributions demonstrate a good spread of candidates across the mark scale in French and Spanish, with this being slightly less so for German.

Figure 6.4: Item-level mark distributions for FN4 for candidates sitting the unit and certificating candidates at A level in the June 2013 exam series

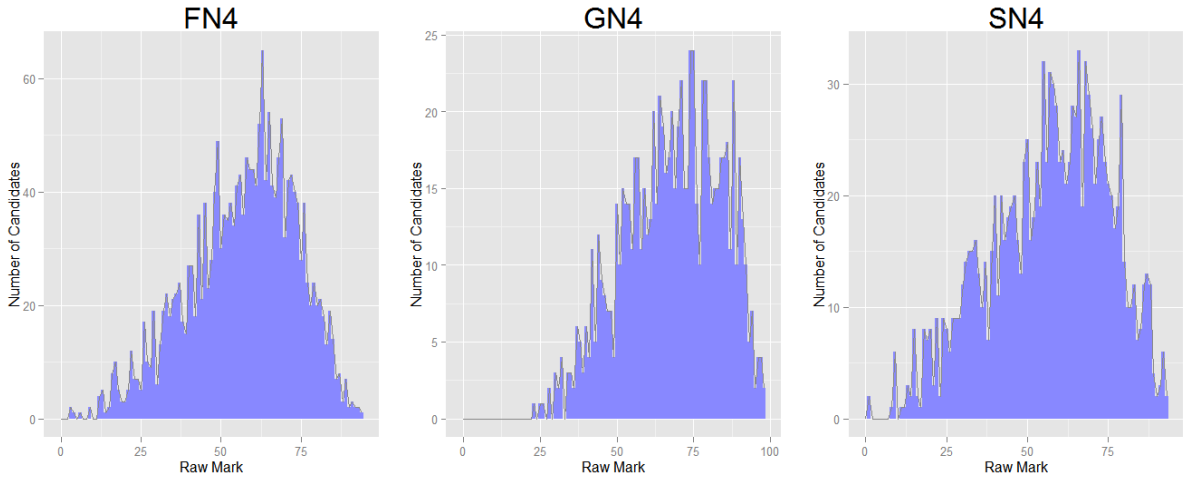


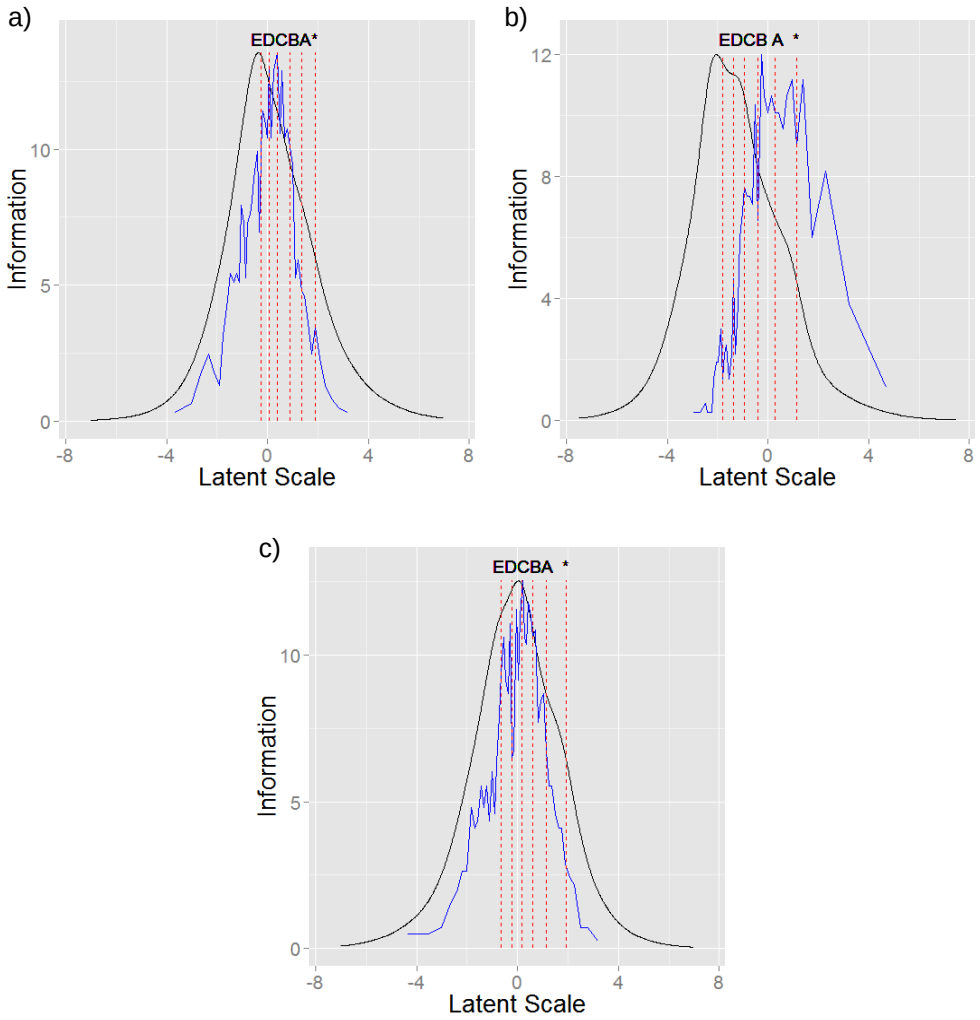
Table 6.3: Descriptive statistics for units FN4, GN4 and SN4 from June 2013 for certificating A level candidates only

Unit	Mean		Standard deviation		Skewness
FN4	56.8	(58.0%)	16.7	(17.0%)	-0.44
GN4	68.0	(69.4%)	15.8	(16.1%)	-0.34
SN4	56.6	(57.8%)	18.5	(18.9%)	-0.39

These analyses have indicated the extent to which the assessments are targeted at the ability of the candidates. However, it does not provide an indication of the effectiveness with which the assessments are targeted at the required standard (that is the grade boundaries). To provide this, the test information functions were estimated using the partial credit model. Due to the high tariff of item 4 on all assessments, fitting these item data using the partial credit model is inappropriate

and therefore these items are excluded from the analysis.[38] The test information functions for units FN4, GN4 and SN4 are presented in Figure 6.5.

Figure 6.5: Test information functions (black) for a) FN4, b) GN4 and c) SN4 from June 2013. Superimposed are the unit-level grade boundaries (dotted) and the distribution of candidate person parameters relative to these information functions (blue).



While FN4 and SN4 are well targeted at the cohort, shown by the overlap between the information function and the distribution of person parameters, there remains a slight discrepancy between the targeting of the assessment and the grade boundaries. A similar relationship between the test information function and the grade boundaries exists for GN4. This suggests it is the higher ability of the German cohort compared to the other subjects that may have given rise to the higher facility

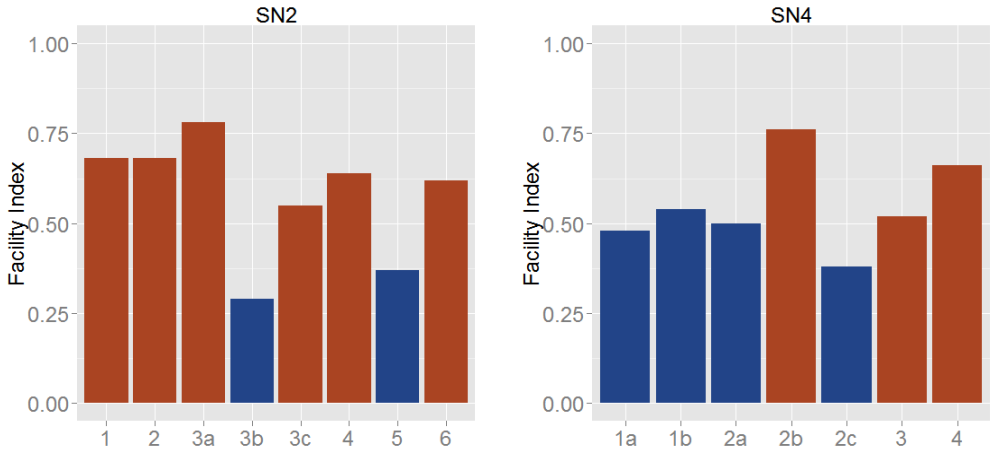
[38] Expected and empirical item-level category probability curves and item characteristic curves are provided in Appendices L, M and N for units FN4, GN4 and SN4, respectively.

indices reported above, rather than a greater discrepancy between the item targeting and the required standard. Increasing the demand of these assessments would support more effective measurement of the ability of the more able candidates.

6.4 Exploration of expert reviewer findings

The expert reviewers raised concerns regarding the clarity of the mark scheme for items 3b and 5 on SN2 and items 1a, 1b, 2a and 2c on SN4. The facility and discrimination indices for these items are presented in Table 6.2, with the facility indices shown graphically in Figure 6.6 to aid visualisation.

Figure 6.6: Facility indices for items on units SN2 and SN4 in June 2013



These figures show that the items flagged by the expert reviewers (blue bars) had the lowest two facility indices on the AS unit and, marginally, four of the lowest five indices on SN4. This, in itself, does not indicate that these items are not functioning correctly and, indeed, the reviewers highlighted the high-order skills involved that may also contribute to the higher demand. However, the approach to defining the mark scheme should be reviewed taking into account the research literature in this area.

Finding 7a indicated that the expectations of candidates' work to achieve marks in the top band appeared very high. However, as shown in Figure 6.3 candidates are achieving marks at the top of the mark scale on all items and the items are discriminating well. This would suggest that either candidates are living up to these expectations in their performances or markers are being appropriately standardised to reflect the assessment context in which responses are being produced.

6.5 Rank ordering of candidates

As described in section 3.7, the correlation between similar measures can be used to indicate potential issues with the validity of a mark distribution. The designation of items to skills for the WJEC AS and A2 written exams is shown in Tables 6.4 and 6.5, respectively.

Table 6.4: Designation of items to skills for the WJEC AS written assessments

June 2012			
FN2 / SN2		GN2	
Item ref	Skill	Item ref	Skill
1	L	1	L
2	L	2	RW
3a	R	3.1	RW
3b	RW	3.2	RW
3c	RW	3.3	RW
4	R	4	RW
5	R	5.1	WC
6	WG	5.2	WC
		6	WG

Table 6.5: Designation of items to skills for the WJEC A2 written assessments

June 2013					
FN4		GN4		SN4	
Item ref	Skill	Item ref	Skill	Item ref	Skill
1	L	1	L	1a	L
2a	RW	2.1	RW	1b	L
2b	RW	2.2	RW	2a	RW
3	WC	2.3	RW	2b	RW
4	WO	2.4	RW	2c	RW
		3	WC	3	WC
		4	WO	4	WO

Figure 6.7 shows AS to A2 intra-skill relationships for the WJEC written exams, with the correlations reported in Table 6.6. The first point of note from Figure 6.7 is the variation in relationships between the different languages, particularly for reading/writing.

The items featuring in the comparison of listening skills are all suitable for objective marking and therefore marking reliability should be very high. The same is true for the reading/writing and reading items where the questions are all seeking to elicit short responses with little subjectivity in the marking. The writing tasks are a combination of extended-response items marked against a levels-of-response mark scheme and a translation task at A2 providing more scope for legitimate variations in marking. Marks for the extended response at A2 total 45, with 15 marks awarded for the quality of response, 10 marks for the knowledge of topics and texts, 10 marks for accuracy and 10 marks for range and idiom. At AS level, the extended response is marked out of 35, with 20 marks available for understanding/quality of response, 10

marks available for accuracy and 5 marks available for range and idiom. The 20 marks awarded for quality of response are not, however, initially awarded on a 20-mark scale. The levels-of-response mark scheme for this aspect of the response is formed from five bands of 2 marks (excluding the separate band for zero marks). This results in a mark out of 10 that is doubled to achieve a mark out of 20. It is likely that this approach is not adequately taking advantage of the available resolution in the mark scale and, therefore, this approach should be revisited with a view to awarding all marks without the need for intra-paper scaling.

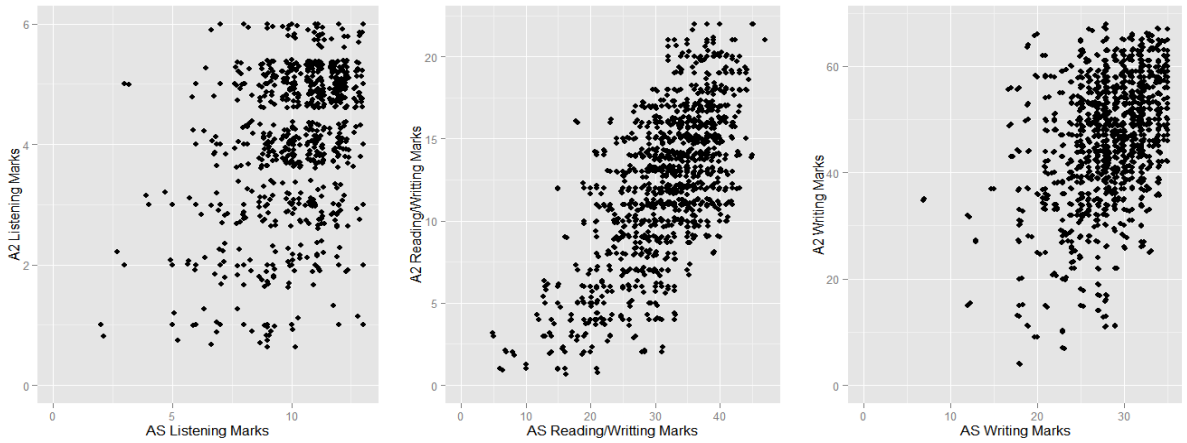
Table 6.6: Correlation between candidates' marks in equivalent skills in AS and A2 assessments for candidates certificating with WJEC at A level in summer 2013

	Skill		Items used for comparison	Item totals	SD – % of item(s) max	A2:AS correlation[39]	No. of candidates
French	Listening	AS	1 2	13	13.5	0.37	762
		A2	1	6	22.9		
	Reading/ Writing	AS	3a 3b 3c 4 5	50	14.1	0.66	768
		A2	2a 2b	22	20.5		
	Writing	AS	6	35	12.4	0.41	766
		A2	3 4	70	18.8		
German	Listening	AS	1	8	16.1	0.49	357
		A2	1	6	23.2		
	Reading/ Writing	AS	2 3.1 3.2 3.3 4	39	13.7	0.74	360
		A2	2.1 2.2 2.3 2.4	22	16.4		
	Writing	AS	5.1 5.2 6	51	13.1	0.75	360
		A2	3 4	70	19.9		
Spanish	Listening	AS	1 2	13	16.6	0.47	503
		A2	1a 1b	6	24.2		
	Reading/ Writing	AS	3a 3b 3c 4 5	50	16.8	0.75	518
		A2	2a 2b 2c	22	20.7		
	Writing	AS	6	51	15.2	0.77	515
		A2	3 4	70	21.6		

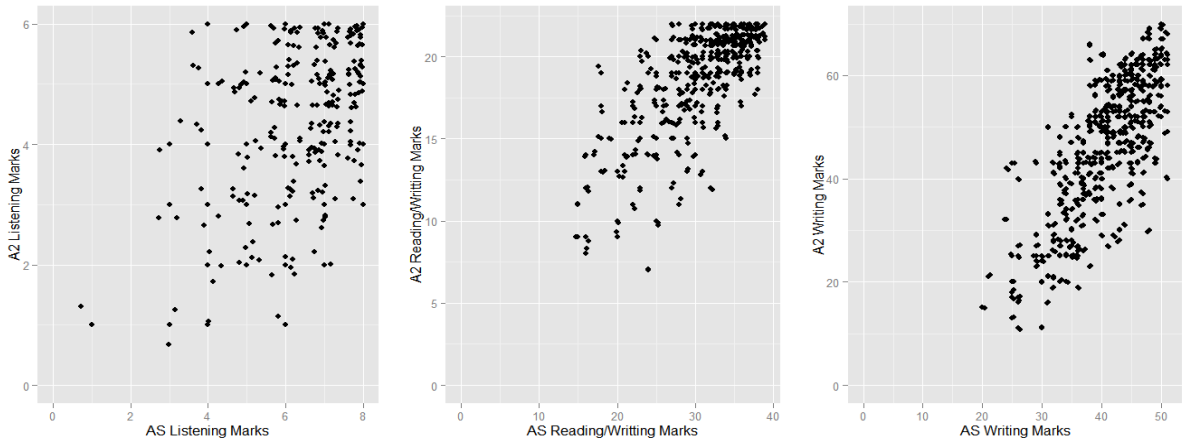
[39] The correlation coefficients are evaluated with candidates removed who scored a total of zero marks for either all AS or all A2 items identified.

Figure 6.7: Scatter plots[40] showing WJEC candidates' marks for listening, reading/writing and writing skills in French, German and Spanish at AS and A2 level for candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

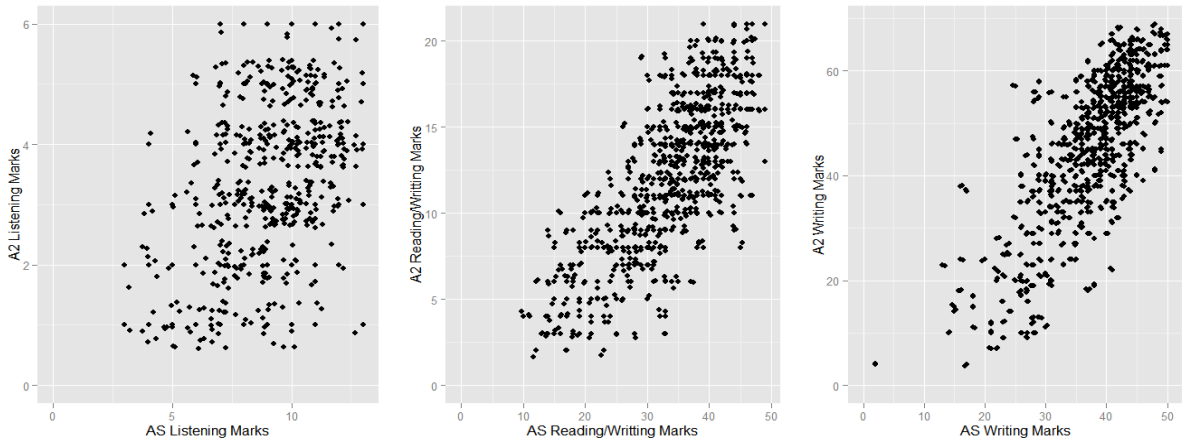
a) French



b) German



c) Spanish



[40] The data points have been jittered to avoid over-plotting to show the density of points.

The first point to note from the intra-skill relationships is the low correlation for listening for all three languages. To an extent, these correlations are low due to the short mark scales associated with this skill. However, given the objective nature of the marking, the correlation is surprisingly low and the functioning of these items should be reviewed.

The correlations for all other skills are satisfactory and appear good for the writing tasks given the subjective nature of the marking, with the exception of the writing skill in French. One finding of the expert reviewers (Appendix D, findings 1a and 2a) was that there are inconsistencies in design approach and therefore consideration should be given to whether such differences in practice may be giving rise to a difference in performance in the assessment of this skill in French.

6.6 Weighting of skills

Shown in Table 6.7 are the intended and achieved weights of the skills assessed in the WJEC A2 written units. This shows that there is a close match between the intended and achieved weight of the skills for French, with some greater differences for Spanish, although they are not of a particularly concerning size. The differences between the intended and achieved weights for German are, however, larger and reflect the reduced discrimination of listening and reading/writing items. This shows that candidates' performances on the writing task have a greater influence on the final rank order than was intended in the design. These values of achieved versus intended weights should be monitored as other recommendations relating to the design of these assessments are delivered.

Table 6.7: Achieved vs intended weight for A2 WJEC written assessments.

Unit	Skill designation	Intended weight %	Achieved weight %	Difference (% pnts)
FN4	L	6.1	5.0	-1.1
	RW	22.4	22.6	+0.2
	W	71.4	72.4	+1.0
GN4	L	6.1	4.8	-1.3
	RW	22.4	14.7	-7.7
	W	71.4	80.5	+9.1
SN4	L	6.1	5.0	-1.1
	RW	22.4	19.4	-3.0
	W	71.4	75.6	+4.2

6.7 Speaking assessments

Figure 6.8 shows the unit-level mark distributions for the WJEC A2 spoken assessments, with the summarising descriptive statistics in Table 6.8. The mean marks awarded on these units are all extremely high, with the distribution becoming compressed and the bottom half of the mark distributions being largely unused. A large number of candidates can also be seen to be achieving the maximum mark possible for the unit. This means that all discrimination between the abilities of these candidates is lost. To explore this further, the relationship between candidates' marks on the speaking assessment at AS (unit 1) and their marks on the equivalent A2 unit (unit 3) was explored. Figure 6.9 shows this relationship for candidates sitting unit 1 in summer 2012, unit 3 in summer 13 and also certificating to A level in summer 2013. The correlation coefficients corresponding to these plots are provided in Table 6.9. Also presented are the modified correlation coefficients as described in section 3.7 due to the considerable ceiling effect present for all subjects.

These modified correlation coefficients are relatively low when compared with the other intra-skill relationships. While it is not possible to determine whether or not this low correlation arises from variations in marking or from an alternative source, the reduced discrimination between high-achieving candidates is likely to contribute. Consideration should be given to addressing this issue.

Figure 6.8: Unit-level raw mark distributions for FN3, GN3 and SN3 for certificating A level candidates in summer 2013 with grade boundaries superimposed

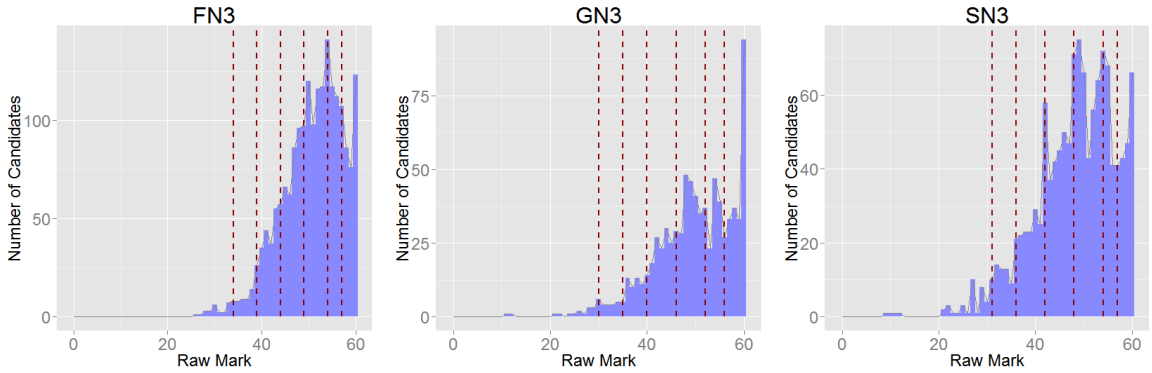


Table 6.8: Descriptive statistics for units FN3, GN3 and SN3 from June 2013 for certificating A level candidates only

Unit	Mean	Standard deviation	Skewness
FN3	50.8 (84.7%)	6.4 (10.7%)	-0.73
GN3	49.6 (82.7%)	8.1 (13.5%)	-0.88
SN3	48.1 (80.2%)	8.3 (13.8%)	-0.87

Figure 6.9: Scatter plots[41] showing candidates' marks for speaking in French, German and Spanish at AS and A2 level for WJEC candidates sitting the written AS exam in summer 2012, the written A2 exam in summer 2013 and certificating at A level in summer 2013

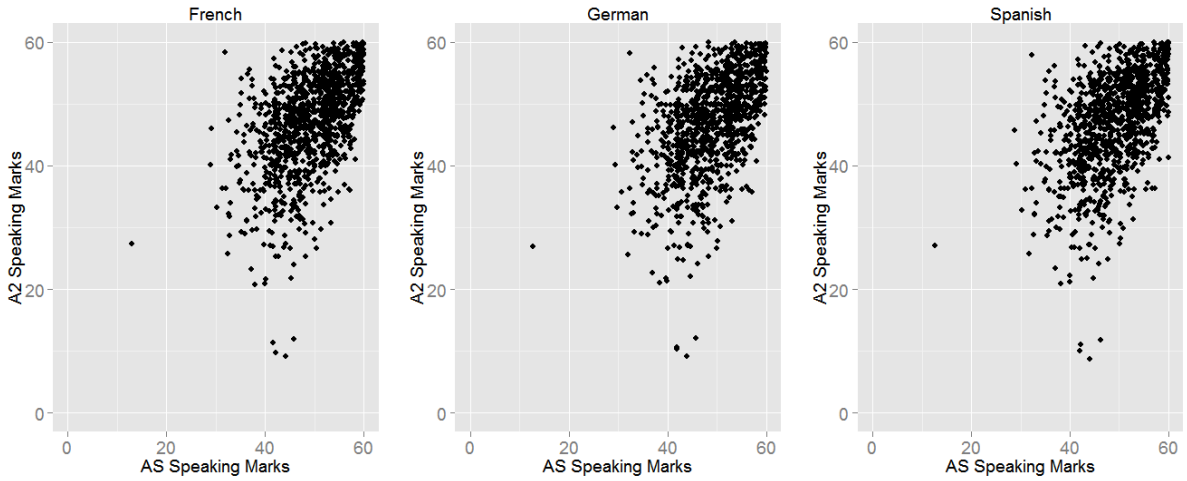


Table 6.9: Correlation between candidates' marks achieved in speaking at AS level (unit 1) and A2 level (unit 3)

	AS:A2 correlation	No. of candidates in correlation	A2:AS modified correlation	No. of candidates in modified correlation
French	0.57	1,802	0.42	1,306
German	0.62	736	0.43	564
Spanish	0.57	1,156	0.43	931

[41] The data points have been jittered to avoid over-plotting to show the density of points.

7 Impact of assessment functioning on A* outcomes

The focus of this report has been the functioning of assessments. However, stakeholders have also questioned A level MFL grading standards. While consideration of the methodology used to award these subjects is out of the scope of this report, and will be picked up by our inter-subject comparability research, the impact of the functioning of the assessments on grading is worthy of consideration. Specifically, considered here is the impact that the shape of the unit-level raw mark distribution has on the position of the unit-level A* conversion point.

The A* conversion point is not a grade boundary as such, insofar as it is it not possible for a candidate to be awarded an A* for an individual unit. This point does, however, indicate the number of raw marks required on an A2 unit for a candidate to be awarded 90 per cent of the available uniform marks for that unit. For a candidate to achieve an A* overall, he/she must achieve 80 per cent of the uniform marks available across the whole of the AS and A2 units combined in addition to achieving at least 90 per cent of the available uniform marks available at A2. Therefore, while achieving above the A* conversion point on any unit does not guarantee a candidate will receive an A* overall (which is also the case for any other unit level grade), candidates scoring above this point are more likely to achieve an A* overall. For ease of reference, candidates are referred to here as achieving an A* at unit level if they have achieved a raw mark at or above the unit-level A* conversion point. However, as explained above, this terminology is not strictly correct.

The A* conversion point is calculated arithmetically. This is in contrast to the grade A and E boundaries, which are judgemental grade boundaries for A level qualifications. These being judgemental grade boundaries means that work on these grade boundaries is scrutinised by senior examiners in awarding meetings and statistical evidence is provided to awarders at these grade boundaries to support their judgements. All other grade boundaries (and conversion points) are calculated arithmetically.[42] The position of the A* conversion point is determined by the following steps:

1. Calculate the difference, in raw marks between the A grade boundary and the B grade boundary.

2. Add this difference to the A grade boundary to determine the provisional A* conversion point.

[42] Further details of the calculation of arithmetic grade boundaries and UMS can be found at www.store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF

3. If this provisional A* conversion point falls halfway between the A grade boundary and the maximum mark or nearer to the grade A boundary than halfway, the provisional A* conversion point stands.

4. However, if the A* conversion point falls nearer to the maximum mark than the A grade boundary, the A* conversion point is placed halfway between the A grade boundary and the maximum mark.[43]

To explore the impact of the shape of mark distribution on unit-level A* outcome, grade boundaries were set on modelled mark distributions. For this purpose, beta binomial distributions with a range of shape parameters ($2 \leq \alpha,\beta \leq 8$, interval = 0.1) and a maximum raw mark of 100 were simulated, giving rise to distributions with a range of means, standard deviations and skews. An example of these modelled distributions is illustrated in Figure 7.1. For the purposes of this modelling, the grade A and E boundaries were set to achieve outcomes as close as possible to 40 per cent and 95 per cent, respectively.[44] The A* conversion point was calculated using standard uniform mark scale (UMS) arithmetic calculation procedures outlined above and the percentage of candidates achieving at or above this point evaluated.

The resulting A* outcomes plotted against skew of the distribution are shown in Figure 7.1a with the equivalent grade A outcomes provided in Figure 7.1b, demonstrating that the statistical standard at grade A was roughly maintained irrespective of the shape of the distribution.

[43] Where there are an odd number of raw marks between the A grade boundary and the maximum mark, the position of the A* conversion point is rounded down when performing this calculation.

[44] These grade A and E outcomes were selected as they are broadly representative of typical outcomes across the A level MFL units.

Figure 7.1: Selection of modelled beta-binomial mark distributions

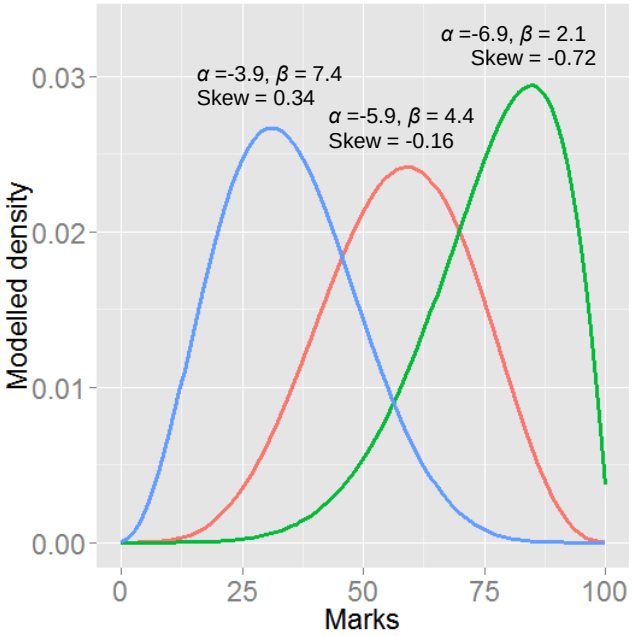
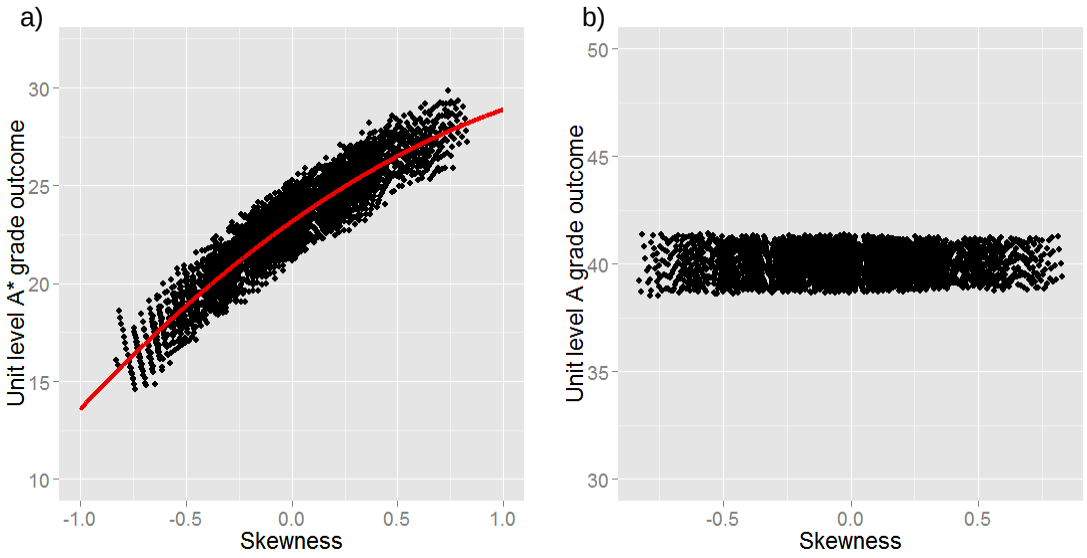


Figure 7.2: a) Grade A* and b) A cumulative percentage unit-level outcomes for modelled units with different levels of skew



It is clear from Figure 7.2a that the skew of the mark distribution has a considerable impact on the outcome at A* even when maintaining the standard at grade A (Figure 7.2b). Mark distributions with more positive skewness are likely to naturally lead to A* outcomes that are higher compared to a more negatively skewed distribution.

Table 7.1: Observed skewness for the summer 2013 A2 written exams for all exam boards and the modelled impact on unit-level A* outcomes of de-skewing those distributions

	Unit	Skewness	Modelled 'de-skewed' change in A* outcome (% pnts)
AQA	FREN3	-0.22	+1.77
	GERM3	-0.63	+5.57
	SPAN3	-0.42	+3.55
OCR	F704	-0.49	+4.20
	F714	-0.51	+4.39
	F724	-0.25	+2.03
Pearson	6FR04	-0.40	+3.36
	6GN04	-0.27	+2.20
	6SP04	-0.49	+4.20
WJEC	FN4	-0.44	+3.73
	GN4	-0.34	+2.81
	SN4	-0.39	+3.27

Table 7.2: Observed skewness for the summer 2013 A2 spoken assessments for all exam boards and the modelled impact on unit-level A* outcomes of de-skewing those distributions

	Unit	Skewness	Modelled 'de-skewed' change in A* outcome (% pnts)
AQA	FREN4	-0.71	+6.39
	GERM4	-0.54	+4.68
	SPAN4	-0.76	+6.91
OCR	F703	0.05	-0.38
	F713	-0.40	+3.36
	F723	-0.31	+2.55
Pearson	6FR03	-0.20	+1.60
	6GN03	-0.43	+3.64
	6SP03	-0.44	+3.73
WJEC	FN3	-0.73	+6.60
	GN3	-0.88	+8.21
	SN3	-0.87	+8.10

To set this in context, in Tables 7.1 and 7.2 are the skewness values (written exams and spoken assessments, respectively) that have been reported throughout this paper for the A2 units.[45] To allow the consequences of the observed skew to be approximately modelled in terms of unit-level A* outcomes, a quadratic function was fitted to the points in Figure 7.2a, as indicated by the red line. Using this expression,[46] the impact at unit level on the A* outcomes if the skew were to be removed from the assessments is also presented in the tables. These values are only indicative since the analysis assumes that this exact relationship is obeyed in in instances and that the modelled A and E grade outcomes (40 per cent and 95 per cent, respectively) are imposed on all units. The approximations are therefore considerable. While it is unrealistic to expect the assessments to be redesigned/modified so that this skew is completely and accurately removed (particularly in the case of the spoken assessments), this shows that even relatively small changes in skew can impact greatly on the unit-level outcome. While this effect may be reduced when aggregated to subject level, removing the negative skew from the mark distributions by increasing the demand is still likely to have a positive impact on A* outcomes.

When the A* grade was introduced at A level for those certificating in summer 2010, statistical tolerances of ±2 per cent from statistical predictions were applied to the A* outcomes. These A* predictions were formed based on applying the current A* calculation rules to the pre-2010 specifications. Therefore, any skewed functioning such as that reported here that may have been present in previous iterations of the specifications may have been carried forward to the current specifications. The motivation for these tolerances was to promote commonality of A* standard between exam boards. This was necessary to protect against the differently shaped mark distributions, that would inevitably result from the assessments delivered by the different exam boards, leading to differences in the A* standard. In instances where A* outcomes were outside of this tolerance, exam boards considered adjusting the A* conversion point(s) from the positions defined above in order to meet these statistical predictions and to protect against inter-exam board differences in standard.

It is important to note that this evidence shows that any intentional or unintentional reduction in the demand of the assessments leading to an increased negative skew of the mark distributions (through marking or otherwise) would act to reduce rather than increase the number of candidates achieving an A* – or, at the very least, may result in outcomes not responding in a manner that may be anticipated.

---

[45] These values are drawn from Tables 3.6, 4.3, 5.3 and 6.3 for the written exams. The skewness values for the speaking assessments are either those values quoted with zero marks removed or are recalculated versions on this basis.

[46] A* outcome = $(-1.93 \times \text{skewness}^2) + (7.64 \times \text{skewness}) + 23.18$.

8 Assessment of cultural aspects

Although implemented differently, all specifications involve the assessment of candidates' wider cultural understanding through the requirement for candidates to, in general, write an extended response focusing on economic, ecological, historical, literary or wider cultural aspects of the country where the language is spoken. Candidates are rewarded for the content of these responses, including (to greater or lesser extents in different exam boards) their ability to demonstrate knowledge of the subject area, the relevance of their response, their personal reaction to the subject area, their ability to deliver a balanced argument, and their imagination and insight.

These aspects of knowledge and understanding do not feature in the current assessment objectives (Table 8.1) and the ability to demonstrate these wider skills appears to be only weakly linked to the skills that the current assessment objectives do articulate. On this basis, the assessment of these content-related aspects would seem to negatively impact on the validity of the assessments, especially given their prominence (in terms of marks) across most writing tasks currently offered.

Table 8.1: Assessment objectives for the current specifications

Assessment objective	Definition
AO1	Understand and respond, in speech and writing, to spoken language
AO2	Understand and respond, in speech and writing, to written language
AO3	Show knowledge of and apply accurately the grammar and syntax prescribed in the specification

It is articulated, however, in the current subject criteria that developing knowledge and skills in this area is important in these specifications. The Aims and Objectives[47] state that the specifications should encourage students to:

> develop awareness and understanding of the contemporary society, cultural background and heritage of countries or communities where the language is spoken;

[47] From www.ofqual.gov.uk/documents/gce-as-and-a-level-subject-criteria-for-modern-foreign-languages-mfl

Inclusion of similar elements has been proposed as part of the current reform activity,[48] the details of which are currently subject to consultation[49] for inclusion in the revised specifications. This consultation proposes "…a new emphasis on the culture and society of the country or countries where the language is spoken, and a requirement for critical analysis and evaluation" and outlines the proposed assessment objectives as found in Table 8.2.

Table 8.2: Assessment objectives as proposed in the consultation for reformed specifications

Assessment objective	Definition
AO1	Understand and respond, in speech and writing, to spoken language drawn from a variety of sources, including face-to-face interaction
AO2	Understand and respond, in speech and writing, to written language drawn from a variety of sources
AO3	Manipulate the language accurately and appropriately, in spoken and written forms, using a range of lexis and structure
AO4	Show knowledge and understanding of the culture and society of countries and communities where the language is spoken and demonstrate critical analysis and evaluation of works created in the language studied

Given these points and the apparent view that this is a set of skills relevant to the domain, provided the items and mark schemes are appropriately developed and delivered, this should not be regarded as a challenge to validity.

[48] Redesigned A level MFL specifications due to be in centres for first teaching in September 2016.
[49] http://comment.ofqual.gov.uk/developing-new-qualifications-for-2016/3-subject-specific-proposals/modern-foreign-languages

9 'Ideal' grade boundary placement and assessment targeting

A common reference when designing or monitoring the functioning of assessments is the location of the grade boundaries and, more specifically, the location of these grade boundaries relative to some 'ideals' or design positions. The motivation for doing this is that it provides an operationally accessible route to considering whether or not the demand of an assessment is targeted appropriately. It is important to note that, regardless of the definition of 'ideal' grade boundaries, the priority must be the maintenance of standards. However, the position of these grade boundaries can be used during the assessment quality/design feedback loop to determine whether or not attempts should be made to target the demand of the assessment differently. This can have technical assessment functioning benefits beyond the maintenance of standards.

A number of (potentially competing) technical definitions or motivations for positioning grade boundaries exist beyond the overriding standards driver, which means that a single definition of these 'ideals' does not exist. One definition used with the current assessments under the UMS system is to attempt to design assessments that result in grade boundaries that lead to a linear conversion from raw marks to uniform marks. Such a relationship would mean that, for each raw mark that candidates achieve, they are awarded a (nominally) identical number of additional uniform marks, regardless of whether they are close to the top, middle or bottom of the mark range. This linear relationship is achieved by grade boundaries that are placed at a proportionally identical position on the raw mark scale to the proportion of the maximum uniform marks available at that grade boundary. For A levels, this means that the 'ideal' A grade and E grade boundaries occur at 80 per cent and 40 per cent of the maximum raw mark, respectively. Table 9.1 shows the judgemental grade boundaries and their position relative to these ideals for the A2 units in summer 2013.

Alternative, and more direct, indicators of test targeting (namely the location of the test information function relative to the grade boundaries) have been used here. Given the evidence presented, this driver for assessment design should be prioritised over linearity of the raw-to-uniform mark conversion.

Table 9.1: Judgemental grade boundary positions relative to one definition of their 'ideal' location

	Unit	Grade	Boundary	Max mark	Difference from 'ideal'
AQA	FREN3	A	86	110	-1.8%
		E	44		+0.0%
	FREN4	A	43	50	+6.0%
		E	25		+10.0%
	GERM3	A	94	110	+5.5%
		E	56		+10.9%
	GERM4	A	43	50	+6.0%
		E	25		+10.0%
	SPAN3	A	91	110	+2.7%
		E	56		+10.9%
	SPAN4	A	43	50	+6.0%
		E	25		+10.0%
OCR	F703	A	46	60	-3.3%
		E	26		+3.3%
	F704	A	103	140	-6.4%
		E	55		-0.7%
	F713	A	47	60	-1.7%
		E	26		+3.3%
	F714	A	112	140	+0.0%
		E	56		+0.0%
	F723	A	47	60	-1.7%
		E	25		+1.7%
	F724	A	100	140	-8.6%
		E	53		-2.1%
Pearson	6FR03	A	37	50	-6.0%
		E	21		+2.0%
	6FR04	A	72	100	-8.0%
		E	37		-3.0%
	6GN03	A	43	50	+6.0%
		E	23		+6.0%
	6GN04	A	78	100	-2.0%
		E	39		-1.0%
	6SP03	A	41	50	+2.0%
		E	22		+4.0%
	6SP04	A	69	100	-11.0%
		E	44		-6.0%
WJEC	FN3	A	54	60	+10.0%

		E	34		+16.7%
	FN4	A	74	98	-4.5%
		E	43		+3.9%
	GN3	A	52	60	+6.7%
		E	30		+10.0%
	GN4	A	80	98	+1.6%
		E	43		+3.9%
	SN3	A	54	60	+10.0%
		E	31		+11.7%
	SN4	A	73	98	-5.5%
		E	36		-3.3%

10 Findings and recommendations

Summarised below are the broad findings resulting from the qualitative and quantitative analysis.

10.1 AQA

- The approach to awarding candidates marks for quality of language in the extended writing tasks does not appear to have a sound basis. The current approach is highly likely to be having a negative impact on the rank order of candidates and therefore the validity of the assessment.
- The prevalence of items that are of relatively low demand for those candidates sitting the written assessment is having a negative impact on the valid discrimination between candidates, especially for the most able candidates.
- The targeting of the written assessments relative to the required standard is suboptimal. This means that there is a greater amount of information collected to differentiate between candidates at the lower-ability range (where there are fewer candidates) than those of higher ability. This is more pronounced for German and Spanish than for French.
- The tendency for the lower-demand items to be concentrated in the assessment of listening for all languages has impacted on the extent to which that skill exerts influence over candidates' final outcomes. This has led to systematic differences between the intended and achieved weighting of skills, with listening being consistently underweighted and writing overweighted.
- The raw mark distributions for the speaking assessments are highly negatively skewed, with a large number of candidates achieving very high marks. This is likely impacting on the discrimination between the more able candidates on these assessments.

10.2 OCR

- Despite many items spreading candidates well across the mark distributions, the written exams contain a high proportion of items with a relatively high facility index, with some offering little discrimination between candidates.
- The targeting of sections A and B of the written exams relative to the required standard is suboptimal. This means that there is a greater amount of information collected to differentiate between candidates at the lower-ability range (where there are fewer candidates) than those of higher ability. This is more pronounced for German than for French and Spanish.

- There appears to be a lack of clarity and principle regarding the definition of acceptable responses for the translation task. Inconsistent principles may impact on the validity of the rank order of candidates.

- The raw mark distributions for the speaking assessments are negatively skewed, with the mark distribution for German containing a large number of candidates achieving maximum marks. This results in a lack of discrimination between the most able candidates.

10.3 Pearson

- The correlations between candidates' reading and writing marks at AS and A2 level are low. This suggests a potentially low level of marking reliability that is impacting on the rank order of candidates and, therefore, validity of the mark distribution.

- A scaling factor of less than one is applied to marks resulting from the translation tasks. This leads to an unnecessary reduction in the discrimination between candidates on this element of the written assessments.

- The raw mark distributions for the speaking assessments are negatively skewed, with truncated distributions at the top of the mark scale. This suggests that the discrimination between candidates at the top end of the ability is reduced.

10.4 WJEC

- The targeting of the combined listening, reading and (compulsory) writing sections of the written exams relative to the required standard is suboptimal, with a greater amount of information collected to differentiate between candidates at the lower-ability range (where there are fewer candidates) than those of higher ability. However, this is extremely marginal for Spanish where the targeting of the exam appears to be broadly appropriate.

- The written assessments in French and Spanish are well targeted to the ability of candidates sitting, with candidates being spread across the mark distribution. This is less so the case for the German exam where there are a number of items with high facility indices meaning they offer little to the discrimination between candidates.

- Even when accounting for the relatively short mark scale, the relationship between AS and A2 marks for listening is weak for all languages. Given that the marking of these items is largely objective, this may suggest issues with item design in this area that require further investigation.

- Marks for the quality of response element of the writing task at AS level are doubled, as the mark scheme has a maximum of 10 marks yet the design is for this element to carry 20 marks. This approach does not, therefore, discriminate between candidates with the resolution that is likely possible when marking this task.

- The raw mark distributions for the speaking assessments are highly negatively skewed, with a large number of candidates achieving very high marks and large regions of the mark distribution being unused. This is likely impacting on the discrimination between the more able candidates on these assessments across all languages.

10.5 Wider findings

- All exam boards, to varying degrees, assess the content of the responses provided in the writing tasks in addition to the quality of the written response. This aspect is not reflected in the assessment objectives for the current specification. Given its inclusion in the Aims and Objectives of the current subject criteria and its proposed inclusion as an assessment objective in the reformed specifications, this is not viewed as compromising the validity of the assessments.

- In an attempt to prevent candidates from being rewarded for pre-prepared responses, a number of mark schemes articulate the manner in which these responses should be credited. These strategies represent a significant risk if the rationale to identifying a pre-prepared response is not clear and justified by evidence. Misidentification or misapplication of an approach would have a negative impact on the rank order of candidates and therefore the validity of the assessment.

10.6 Recommendations

Given these findings, summarised below are the recommendations from this report and the organisation to which those recommendations are relevant. Those marked with * should be considered by exam boards to be required actions that will be followed up by Ofqual.

	Recommendation	Organisation	Justification
*1	The demand of the written assessments must be reviewed in line with the evidence presented in this report. It is strongly recommended that the demand be increased to facilitate more effective measurement of the abilities of the more able candidates. Exam boards must report to Ofqual their approach to addressing this for the assessments to be delivered from summer 2015, along with an action plan and rationale for their approach.	AQA OCR WJEC	Suboptimal targeting of assessment demand relative to the required standard. A high prevalence of items that are relatively low demand for the cohort. Systematic differences between intended and achieved weight of skills.
*2	Consideration must be given to how the assessments (and supporting processes such as standardisation and moderation) of spoken language can be better designed to address the issue of poor discrimination between candidates. It is not expected that spoken language assessments/arrangements are modified from summer 2015, however, opportunities must be sought to improve these assessments in the lifetime of the current specifications in addition to considering alternative approaches in the reformed specifications. Exam boards' reviews and action plans in relation to the current specifications will be followed up.	AQA OCR Pearson WJEC	Raw mark distributions with high mean marks and negative skew in addition to unused parts of the mark scale and truncation of the distribution for high-ability candidates.

3	Consideration must be given to how the assessments (and supporting processes) of spoken language can be better designed in the reformed specifications to improve, monitor and intervene in the quality of marking/consistency of marking standard.	AQA OCR Pearson WJEC	Low correlations suggesting low quality of marking and/or poor discrimination between candidates.
*4	The rationale for capping candidates' quality of language marks in the writing task based on marks achieved for content must be revisited and appropriate modifications to the approach made for the summer 2015 assessments.	AQA	Distorted item-level mark distributions and misapplication of marking rules affecting the rank order on invalid grounds.
*5	Further exploration of additional operational data and assessment/mark scheme design must be performed to understand the low correlation between writing marks, which suggest unsatisfactory item design or quality of marking.	AQA Pearson	Low writing intra-skill correlation.
*6	Further exploration of additional operational data and assessment/mark scheme design must be performed to understand the low correlation between listening marks, which suggests unsatisfactory item design or quality of marking.	WJEC	Low listening intra-skill correlation.
*7	The application of a scaling factor less than 1 to marks from the translation task should be revisited and alternative approaches sought in time for the 2015 assessments.	Pearson	Loss of discrimination through scaling factor.
*8	The approach to up-scaling quality of response marks (10 marks x 2) rather than applying a mark scheme with a sufficient length (20 marks) must be reviewed and addressed in time for the 2015 assessments.	WJEC	Potential loss of resolution in the mark scale.

9	The absence of cultural aspects of knowledge and understanding from the assessment objectives should be considered in the criteria for the reformed specifications as part of the on-going consultation process.	Ofqual/ALCAB	Evidence that these elements are valued as relevant areas of understanding.
*10	The principles underlying the design of the mark scheme and determination of what constitutes an acceptable response must be reviewed for the 2015 assessments and the principles clearly articulated. This will support transparency and future item development.	AQA OCR Pearson WJEC	AQA: expert review finding 2a. OCR: expert review findings 2c, 5a. Pearson: expert review finding 1a. WJEC: expert review findings 2b, 2d.
*11	The principles for defining and crediting pre-prepared responses and targeted lifts from resources must be clarified and articulated for the 2015 assessments reflecting on the findings of the expert reviewers.	OCR Pearson WJEC	OCR: expert review finding 2a. Pearson: expert review finding 2a. WJEC: expert review finding 2c.
*12	Exam boards must monitor the impact of making modifications to the assessments considered here using appropriate metrics as a basis for reporting to Ofqual. Processes should also be put in place for the on-going monitoring of assessment functioning/quality.	AQA OCR Pearson WJEC	Impact of any modifications is necessary for monitoring purposes. On-going good practice in assessment quality monitoring.
13	The principles and practice of handling word limits must be reviewed, clearly articulated and evidence based.	Pearson WJEC	Pearson: expert review findings 1b, 4a. WJEC: expert review finding 2a.
*14	The design of levels-of-response mark schemes must be reviewed including consideration of the comments of the expert reviewers to achieve consistent application of best practice across all languages/mark schemes/optional questions. This must be considered for	AQA OCR Pearson WJEC	AQA: expert review findings 1a, 1b, 1c, 1d, 3c, 4c, 4e, 6a, 6b. OCR: expert review findings 1a, 1b, 2b, 3a, 4a, 6a.

	the written assessments in time for the 2015 assessments.		Pearson: expert review findings 3a, 3c, 3d, 6a. WJEC: expert review findings 1a, 3a.
*15	The comparability of the different optional routes through the assessment must be reviewed in light of the qualitative findings. This must be performed ready for the assessments to be delivered in summer 2015.	OCR	OCR: expert review findings 9a.

10.7 Implications of findings and recommendations for teaching and learning

Increasing the demand of the assessments in line with the recommendations outlined above will improve the validity of the rank order of candidates. There will likely be implications for teaching and learning and the perceptions of users, however, as no change to content or approach is being proposed, modification of what candidates are taught or how they are prepared for exams is not required. It is unlikely that the changes in demand required to effect an improvement in the validity of the assessments will be substantial. However, consideration should be given to how to provide support in these circumstances. While some of these recommendations may appear to have the potential to impact on the grades of candidates, awarding will account for any increase in demand, therefore protecting outcomes.

Appendix A

AQA: expert reviewer findings

Finding of reviewers	Action
1. Clarity of instructions	
a. Some valuable guidance appears only in the *Advice to Teachers* booklet and not in the mark schemes. For example, for unit 1, the advice booklet states that "There is a list of suggested content points for the guidance of examiners but these are by no means prescriptive and students will get credit for well-argued points not in the list" and "There is no mathematical guide to Content marks…."	This issue may influence the findings of the analysis presented in section 3.6. Finding of reviewers should be considered by exam board.
b. For unit 1, section B, writing, the explanation of "Choosing the band for Content" is helpful guidance to markers deciding which mark band candidates' responses should be placed in. However, no instructions are given for awarding marks within each band.	This issue may influence the findings of the analysis presented in section 3.6. Design of the mark scheme for this part of the assessment should be reviewed by the exam board in light of this finding.
c. For unit 1, section B, writing, there are no instructions to markers on how to deal with responses that do not meet the minimum requirement of 200 words.	This issue may influence the findings of the analysis presented in section 3.6. Finding of reviewers should be considered by exam board.
d. For units 2 and 4, speaking, no instructions are given for awarding marks within a band.	This issue may influence the findings of the analysis presented in section 3.6. Finding of reviewers should be considered by exam board.

e. For unit 3, section B, writing, detailed guidance is given on how to apply the mark scheme, taking into account the variation that markers may see in candidates' responses.	–
2. Clarity of principles for crediting candidates' work	
a. For SPAN1, questions 1c and 1d, the principles behind individual items that appear in the "reject" column for discrete answers are not clear. 1c – the correct answer is "unpleasant/not pleasant/not nice". The answer "disagreeable" or "awful" is rejected. 1d – the correct answer is "excessive drinking (alcohol)". The answer "drinking excessive alcohol" is rejected.	This finding is investigated further in section 3.5. Finding of reviewers should be considered further by the exam board as indicated in the summary of recommendations.
3. Sufficient indicative content/terminology	
a. For the unit 1 essay titles, there is some helpful indicative content for all three languages.	–
b. For the unit 3 essay titles in French, the mark scheme provides guidance on the content for individual essays, but this guidance is not provided for German or Spanish.	This finding may impact on inter-subject differences presented in section 3.6. Consistent application of best practice should be reviewed by the exam board.
c. For the writing sections in unit 1, the distinction between some mark band descriptors is difficult to make when there is no further guidance about the interpretation of these terms in the mark scheme. For example, the accuracy marking grid has "Largely accurate with some basic errors" for band 4 and "Generally accurate but still with some basic errors" for band 3. The range of structures grid has "Very good variety of grammatical structures used" for band 5 and "Good variety of grammatical structures used" for band 4.	This issue may influence the findings of the analysis presented in section 3.6. Finding of reviewers should be considered by exam board.
4. Appropriateness and fairness of mark schemes for crediting candidates for what they know, understand and can do	
a. For units 1 and 3, section B, writing, the marks awarded for quality of language (represented by	This finding is explored in section 3.3 and may

three mark grids for range of vocabulary, range of structures and accuracy) cannot be more than one band higher than the band awarded for content. This means that there is the potential for candidates' marks to be reduced three times.	also impact on the findings of the analysis presented in section 3.6. This finding should be considered further by the exam board in line with the recommendation outlined in section 10.
b. For units 2 and 4, part 2 conversation, the marks awarded for interaction are reduced by one band if candidates do not spend the allotted time on each of the topics, which may lead to an inconsistent application of the mark scheme if examiners are focused on timing rather than the quality of response from candidates.	The issues raised may impact on the findings of the analysis presented in section 3.8. This finding should be considered further by the exam board.
c. It is not always clear where a mark of zero should be given for work not worthy of any credit as zero is included in some of the marking bands. For example, in unit 4, speaking, knowledge of grammar grid, there is a mark band of zero to three marks for a performance with the following characteristics: "Generally comprehensible to a native speaker. Limited range of constructions, vocabulary and sentence patterns. Serious grammatical errors may sometimes cause difficulties for immediate comprehension."	This finding should be considered further by the exam board.
d. There is good practice in the unit 3 mark schemes, which state that all work is marked and it is the quality of the response and not the number of words that is important.	–
e. For GERM1, section B, writing, there is a further application of limiting factors, but this is not consistent across the three questions. In question 10, reference is made to a proposed film club ("*Sie… möchten mit Freunden einen Kino-Klub organisieren*") and responses that focus exclusively on a club that is already operating cannot score more than 12/20 for content. In question 11, if candidates do not address the second part of the question they can also only score 12/20. However, in question 12 there is no such limiting statement.	The currently provided data do not allow quantitative examination of this finding. The exam board should investigate this issue further using internally available data sources to establish any impact on inter-route comparability.

5. Points-based mark schemes	
a. For SPAN1, questions 1g and 4c, it is not clear whether all the information in the mark scheme boxes is required or whether these are alternative answers. 1g – gives three acceptable answers, but the question is only worth two marks. 4c – gives two acceptable answers, but the question is only worth one mark.	The finding is considered in section 3.5. This finding should be considered further by the exam board.
6. Clarity in relation to qualities worthy of the higher marks	
a. For units 1, section B, writing, the marking criteria use phrases such as “Wide range of appropriate vocabulary” and “A range of appropriate vocabulary”, which without detailed exemplification are very open to interpretation and there is the potential for inconsistent application.	The issues raised may impact on the findings of the analysis presented in section 3.6. The exam board should consider the extent to which this finding is addressed through the examiner standardisation process.
b. For unit 2, speaking, more amplification of the criteria used to distinguish between good and fairly good work for pronunciation and intonation would be helpful.	The issues raised may impact on the findings of the analysis presented in section 3.8. This finding should be considered further by the exam board.
7. Only perfect answers for top marks?	
a. There are some instances in the mark schemes where top mark bands appear to set very high performance expectations. For example: For unit 4, part 1 discussion of stimulus card, the top band for “In the face of challenges by the examiner” has “Responds readily to all opportunities to develop views and defend and justify opinions.” For unit 4, part 2 conversation, the top band on the	This finding is considered in section 3.8. No further action required by the exam board.

fluency grid describes “A thoroughly confident speaker. Able to sustain a conversation at natural pace.” The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for an 18-year-old candidate should look like at the very highest level, particularly in areas such as fluency.	
8. Advantage or disadvantage for native or non-native speakers?	
a. For unit 4, speaking, the very high expectations of some of the top mark bands may advantage native speakers if the understanding of what a top performance from a non-native-speaking 18-year-old candidate looks like is not consistent.	This finding is considered in section 3.8. No further action required by the exam board.
9. Comparability of different writing tasks	
No issues of note for French, German or Spanish.	–
10. Comparability of different speaking stimuli	
a. For SPAN2, speaking stimulus card B gives more language support than the other cards – five speech bubbles with conjugated verbs that could provide useful material for candidates in their responses, whereas other cards have no conjugated verbs or ones that are less obviously useful to candidates.	The currently available data do not allow investigation of this issue. Further consideration of this finding should be given by the exam board.

Appendix B

OCR: expert reviewer findings

Finding of reviewers	Action
1. Clarity of instructions	
a. Although the mark schemes have essentially the same content, there are differences in the presentation of the mark schemes between French, German and Spanish, with the German booklets appearing more concise and user-friendly than for the other two languages. The mark scheme booklets for German are laid out in landscape style, the font is bigger and marking grids are printed over fewer pages. For example, for unit 2, the German mark scheme has 21 pages, Spanish has 25 pages and French has 29 pages.	This issue may influence the findings of the analysis presented in section 4.5. Finding of reviewers should be considered by exam board. Consistent application of best practice should be reviewed by the exam board.
b. For units 1 and 3, topic discussion speaking tasks, markers are instructed to cap marks for ideas, opinions and relevance (grid D for unit 1 and grid M for unit 3) at four marks for insufficient reference to the target-language country, but does not exemplify what insufficient means.	The currently available data do not allow investigation of this issue. Further consideration of this finding should be given by the exam board.
2. Clarity of principles for crediting candidates' work	
a. For units1 and 3, topic discussion speaking tasks, markers are instructed to put a cap on marks for fluency, spontaneity and responsiveness (grid E.1 for unit 1 and E.2 for unit 3) of four and two marks, respectively, for pre-learned non-spontaneous material, but the principles for judging this need to be made clear.	The currently available data do not allow investigation of this issue. Further consideration of this finding should be given by the exam board.
b. For unit 2, task 7, there is a recommendation to write 200 to 300 words, but no guidance as to how to mark overly long responses.	The issues raised may impact on the findings of the analysis presented

	in section 4.5. This finding should be considered further by the exam board, such as whether this is sufficiently addressed during examiner standardisation.
c. For F724, task 7, translation into English, the principle of having “night-time protest” as an acceptable answer but not allowing “night protest” is not clear, particularly considering this is a transfer-of-meaning exercise.	This finding is explored in section 4.4 and may also impact on the findings of the analysis presented in section 4.5. This finding should be considered further by the exam board in line with the recommendation outlined in section 10.
3. Sufficient indicative content/terminology	
a. For unit 4, essay questions, there is scope for some general indication of content to be given, although the broad, open-ended nature of the titles makes this more challenging.	The issues raised may impact on the findings of the analysis presented in section 4.5. This finding should be considered further by the exam board.
b. For unit 1, section A role plays, there is a good level of indicative content, whereas for unit 3, section A discussion of an article, there is no indicative content	The currently available data do not allow investigation of this issue. Further consideration of this finding should be given by the exam board.

4. Appropriateness and fairness of mark schemes for crediting candidates for what they know, understand and can do	
a. For unit 2, task 7b, the top band for response to text (grid J) is five marks wide, whereas all the other bands are four marks wide. The quality of marking literature review identified that it is good practice to have marks evenly distributed across bands in levels of response schemes.	This finding should be considered further by the exam board.
5. Points-based mark schemes	
a. For F722, task 3e gives two points of information in the answer box – "wide range" and "reasonably priced" – but is only worth one mark.	This finding is explored in section 4.4. This finding should be considered further by the exam board.
6. Clarity in relation to work worthy of higher marks	
a. Terminology is generally consistent, but greater definition of top mark band descriptors would facilitate consistency of interpretation. For example, in unit 4, accuracy of language grid (C.1), the top band indicates the use of complex structures, but does not exemplify what these are for A level.	The issues raised may impact on the findings of the analysis presented in section 4.5. This finding should be considered further by the exam board.
7. Only perfect answers for top marks?	
a. For all units, mark schemes allow for the presence of errors.	–
b. For unit 3, topic conversation, the top band for fluency, spontaneity and responsiveness (grid E.2) has "Responds promptly and fully. Consistently shows initiative. Leads the conversation. A fluent and spontaneous performance throughout." This appears to set very high expectations. The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for an 18-year-old candidate should look like at the very highest level, particularly in areas such as fluency.	This finding is considered in section 4.6. No further action required by the exam board.
8. Advantage/disadvantage for native/non-native speakers?	
a. For unit 3, the very high expectations of the top mark band in fluency of the conversation may advantage native speakers if the understanding of	This finding is considered in section 4.6. No

what a top performance from a non-native-speaking 18-year-old candidate looks like is not consistent.	further action required by the exam board.
9. Comparability of different writing tasks	
a. For unit 4, the nature of the second essay option in each topic area does not appear to be comparable to the first essay option in each topic area. The first essay option is a general discursive essay that leads well to analysis, development of argument and the drawing of conclusions, for example, in Spanish, question 17, “Many people believe that literature and the arts are a good reflection of the society that produces them. Referring to one or two literary or artistic works that you have studied, how do these help you to understand the country?” The second essay option, question 18, requires candidates to write a letter to a cousin recommending what he/she should study at university, which may not elicit the same level of analysis and evaluation as the first option. There are other instances for the second option where candidates are required to write a report or a blog, which does not seem comparable to the traditional discursive essay.	The currently available data do not allow investigation of this issue, but may impact on the analysis in section 4.5. Further consideration of this finding should be given by the exam board.
10. Comparability of different speaking stimuli	
No issues of note for French, German or Spanish.	–

Appendix C

Pearson: expert reviewer findings

Finding of reviewers	Action
1. Clarity of instructions	
a. There are no instructions on how to award marks within bands. This is a particular issue for the wider mark bands – for example, in unit 1 where mark bands in the response grid are four marks wide and in unit 4 for the research-based essay where mark bands in the reading, research and understanding grid are six marks wide. This will not facilitate an accurate and consistent application.	The issue raised may impact on the findings of the analysis presented in section 5.4. This finding should be considered further by the exam board.
b. For unit 4, there are no instructions to markers on how to deal with responses that infringe the word limits for the essay questions 2 and 3.	The issue raised may impact on the findings of the analysis presented in section 5.4. This finding should be considered further by the exam board.
2. Clarity of principles for crediting candidates' work	
a. For unit 2, questions 4 and 7, the mark scheme states that targeted lifts are acceptable, but more detail is needed about the level of lifting that is acceptable so that this is dealt with consistently by markers.	The issue raised may impact on the findings of the analysis presented in section 5.5. This finding should be considered further by the exam board.
3. Sufficient indicative content/terminology	
a.For unit 1, there is no indicative content for the stimulus tasks.	The issue raised may impact on the findings of the analysis presented in section 5.6. This finding should be considered further by the exam board.
b. For unit 4, there is some indicative content for the creative and discursive essays.	–

c. For unit 4, research-based essay questions, indicative content is included for each of the three languages, but the amount provided is inconsistent. There is more for German than for either of French or Spanish.	The issue raised may impact on the findings of the analysis presented in section 5.4. Consistent application of best practice should be reviewed by the exam board.
d. Some of the terminology used in band descriptors is very broad and holistic, which creates the potential for wide variations in interpretation and application. For example, in unit 4, creative and discursive essays, organisation and development grid, band 4 to 6 has “Limited organisation and development not always logical and clear. Structure lacks coherence”, which is very similar to band 7 to 9: “Organisation and development not always logical and clear.” The band descriptors for the research-based essays, reading research and understanding grid are also very brief and this has the highest mark allocation of the specification (30 marks).	The issues raised may impact on the findings of the analysis presented in section 5.4. This finding should be considered further by the exam board.
4. Appropriateness and fairness of mark schemes for crediting candidates for what they know, understand and can do	
a. For unit 2, question 8, the approach to word count could be unfair in the way that it is applied. Markers are instructed to cap marks for ‘content and response’ at 9 out of 15 if there is a missing bullet point, even if the fourth and final bullet point is included, but goes beyond the 220 word limit.	The issue raised may impact on the findings of the analysis presented in section 5.4. This finding should be considered further by the exam board.
5. Points-based mark schemes	
No issues for French, German or Spanish.	
6. Clarity in relation to work worthy of higher marks	
a. There are examples where it is difficult to distinguish clearly between the top two marks bands: in unit 4, creative and discursive essays, range and application of language grid, the top band has “Rich and complex language; very successful	The issues raised may impact on the findings of the analysis presented in section 5.4. This

manipulation of language", which is very similar to the next band down: "A wide range of lexis and structures, successful manipulation of language." For the research essays, reading, research and understanding grid, the top band has "Very good to excellent understanding. Clear evidence of extensive and in-depth reading and research" and the next band down has "Good to very good understanding. Clear evidence of in-depth reading and research."	finding should be considered further by the exam board.
7. Only perfect answers for top marks?	
a. Some of the phrases used in band descriptors for the essay questions in units 6SP02 and 6SP04 appear to set very high expectations. For example, 6SP02, question 8, the top band for the content and response grid has "Task fully grasped, answer wholly relevant…." The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for a 17- or 18-year-old candidate should look like at the very highest level.	This finding is considered in section 5.3. No further action required by the exam board.
8. Advantage or disadvantage for native or non-native speakers?	
a. For units 2 and 4, the very high expectations generated by the phrasing used in some of the top mark bands for the essay questions may advantage native speakers if the understanding of what a top performance from a non-native-speaking 17- or 18-year-old candidate looks like is not consistent.	This finding is considered in section 5.3. No further action required by the exam board.
9. Comparability of different writing tasks	
No issues of note for French, German or Spanish.	
10. Comparability of different speaking stimuli	
No issues of note for French, German or Spanish.	

Appendix D

WJEC: expert reviewer findings

Finding of reviewers	Action
1. Clarity of instructions	
a. There are a number of inconsistencies across the 2013 mark scheme booklets for French, German and Spanish. The French booklet is the most comprehensive, containing the mark schemes for the speaking assessments and the written papers as well as a page of marking principles for AS paper 2 (FN2). The Spanish and German booklets do not contain the marking grids for the speaking assessments or any overall principles for applying the mark schemes. The French and Spanish booklets present both the questions and mark schemes together for parts of the written papers, whereas the German booklet does not do this. The German mark scheme booklet is the least detailed of the three languages. Although band descriptors themselves are detailed, none of the mark schemes contain any instructions on how to choose a mark band and a mark within a band.	The issues raised may impact on the findings of the analysis presented in section 6.5. Consistent application of best practice should be reviewed by the exam board.
2. Clarity of principles for crediting candidates' work	
a. There is no indication of the marking principles if the word limits are infringed for writing tasks or if the duration of the speaking assessments fall short of the required timings. However, for French, there are some separate guidance notes for the AS essays indicating that any work that exceeds 250 words will be crossed out and not marked. This instruction is not included in the mark scheme.	The issue raised may impact on the findings of the analysis presented in section 6.5. Consistent application of best practice should be reviewed by the exam board.
b. The mark schemes for several tasks do not include details of what alternative answers are acceptable and what will be rejected: SN2, reading tasks 3b and 5; SN4, listening tasks 1a and 1b; reading tasks 2a and 2c.	This finding is considered in section 6.4. This finding should be included in line with the recommendation

SN4 reading tasks 2a and 2c in particular require a high level of manipulation, inference, deduction and personal opinion, so markers need to be clear what responses are acceptable and what should be rejected.	outlined in section 10.
c. In SN2, reading task 3b and SN4, reading task 2c, candidates are instructed to use their own words, but there is no indication of the principles applied if they do not do this. The same is true for FN2, reading task 3b, but the principles on page 7 indicate that "No marks will be awarded for copying from the text in most cases."	The issues raised may impact on the findings of the analysis presented in section 6.4. Further consideration of this finding should be given by the exam board.
d. For SN4, task 3, translation into Spanish, there is no indication of alternative acceptable responses or those that should be rejected to show markers what is worthy of credit. Also, markers are instructed to refer to the published grid for accuracy marks, but this is not included next to the task in the mark scheme, whereas for French it is. This grid is also missing from task 3 in the German mark scheme.	The issues raised may impact on the findings of the analysis presented in section 6.4. Further consideration of this finding should be given by the exam board.
3. Sufficient indicative content/terminology	
a. There is no indicative content for any of the essay questions at AS or A2, although at A2 there are between 48 and 54 individual options for the guided studies component.	The issues raised may impact on the findings of the analysis presented in section 6.4. Further consideration of this finding should be given by the exam board.
4. Appropriateness and fairness of mark schemes for crediting candidates for what they know, understand and can do	
a. For unit 1, speaking, communication grid and unit 2, essays, understanding/quality of response grid, the marks that candidates are awarded out of 10 is doubled for a total out of 20, which means that	Further consideration of this finding should be given by the exam

candidates can only achieve 'even' marks.	board.
5. Points-based mark schemes	
No issues of note for French, German or Spanish.	
6. Clarity in relation to work worthy of higher marks	
No issues of note for French, German or Spanish.	
7. Only perfect answers for top marks?	
a. For SN4, some of the words/phrases used in band descriptors appear to set very high expectations. For example, for range and idiom, the top band describes "Assured sense of register. Uses language imaginatively to achieve desired effect. Evidence of style, nuance…." The key issue is interpretation of the descriptors and markers having a common understanding of what the performance standard for an 18-year-old candidate should look like at the very highest level.	The issues raised may impact on the findings of the analysis presented in section 6.4. No further action.
8. Advantage or disadvantage for native or non-native speakers?	
No issues of note for French, German or Spanish.	
9. Comparability of different writing tasks	
No issues of note for French, German or Spanish.	
10. Comparability of different speaking stimuli	
No issues of note for French, German or Spanish.	

Appendix E

Key to skills abbreviations

L = Listening

R = Reading (short-answer items)

RW = Reading/Writing (for marks that cannot be distinguished between reading or writing, such as translation items (both into and out of target language) or extended written-response reading comprehension items)

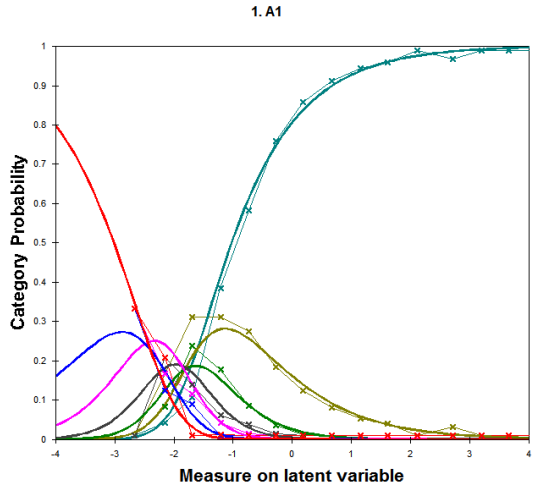
WC = Compulsory extended-response open-writing item

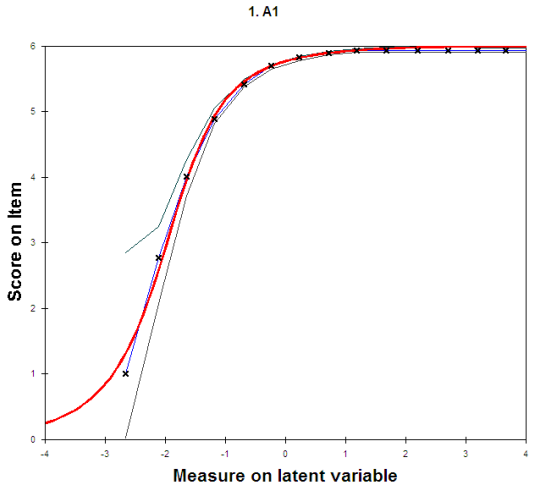
WO = Optional extended-response open-writing item

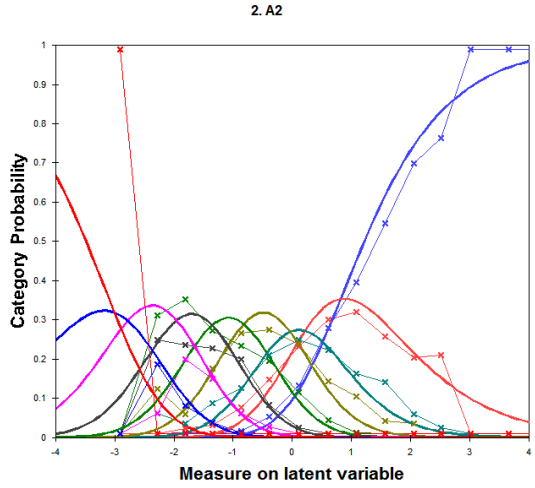
WG = Writing general (for overarching marks or marks not broken down into compulsory/optional items)

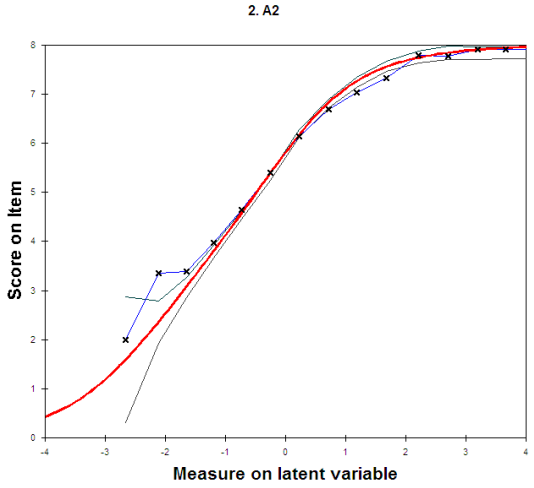
Appendix F

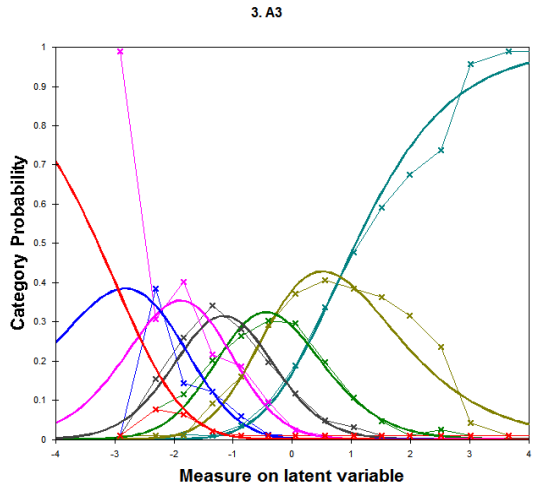
Item-level fit characteristics for AQA's FREN3 in June 2013

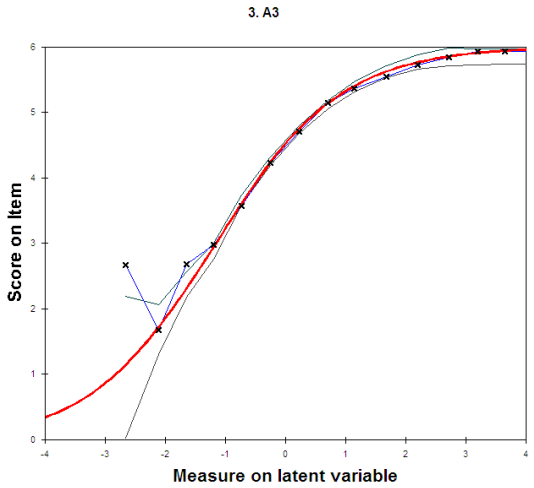


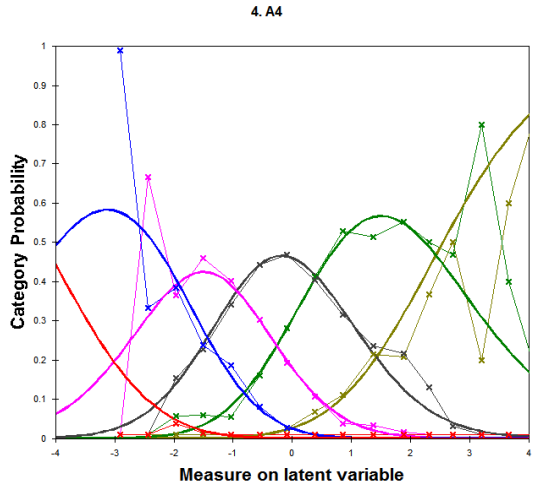


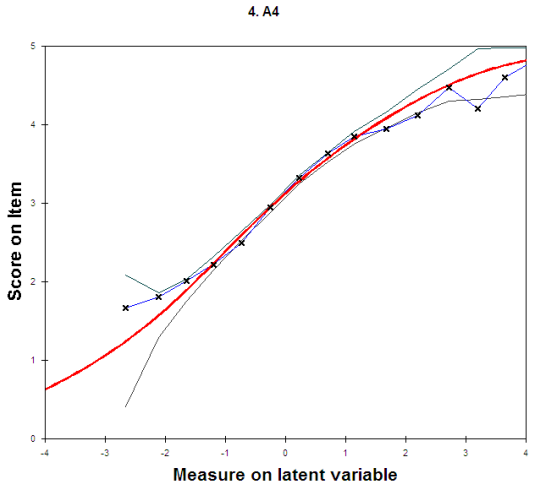


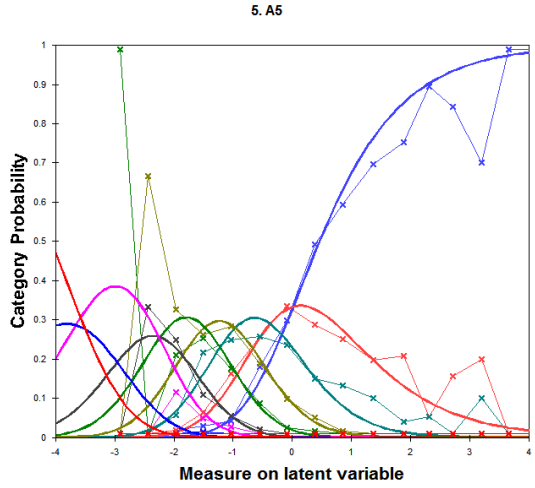


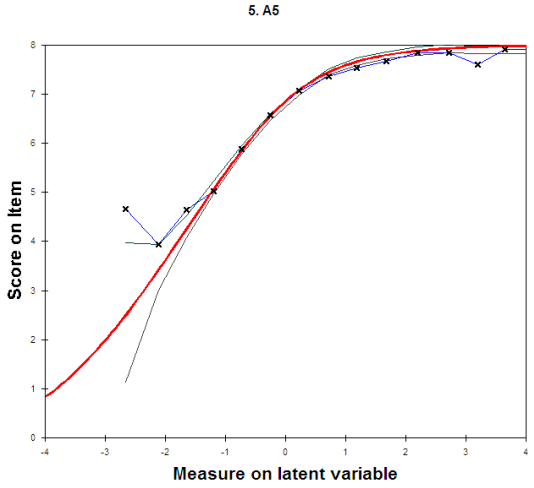


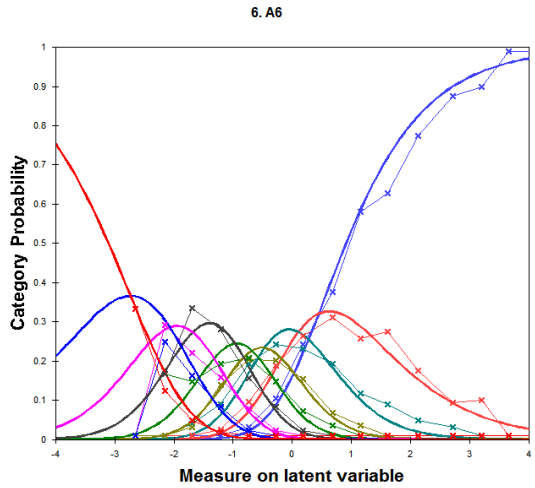


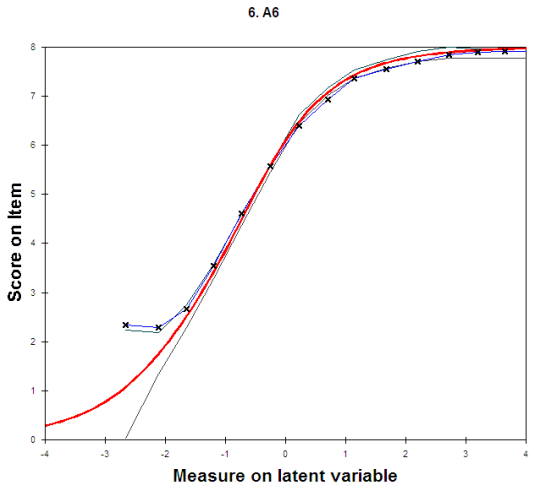

4. A4
Category Probability
Measure on latent variable

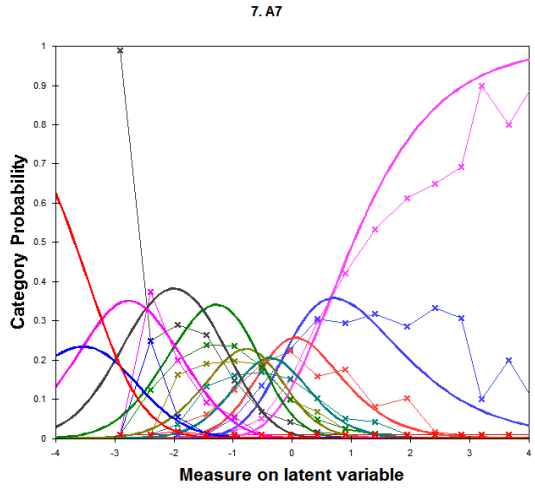

4. A4
Score on Item
Measure on latent variable

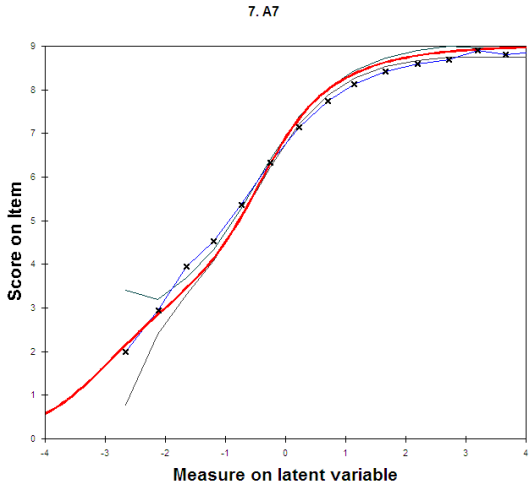

5. A5
Category Probability
Measure on latent variable

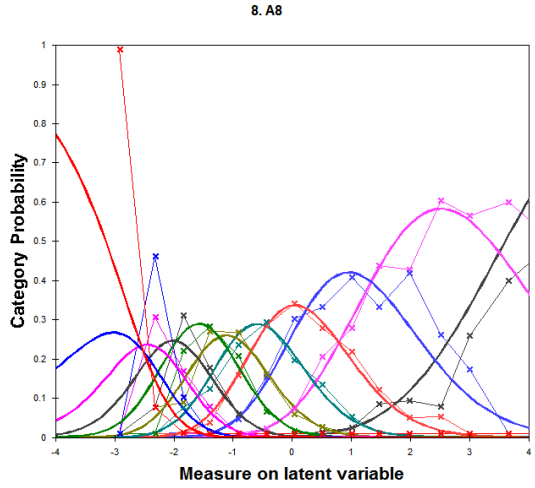

5. A5
Score on Item
Measure on latent variable

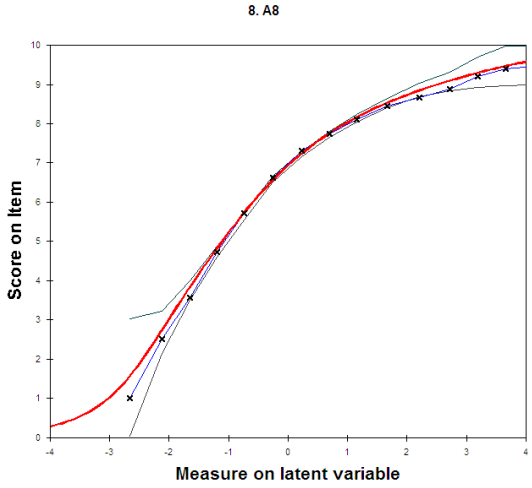

6. A6
Category Probability
Measure on latent variable

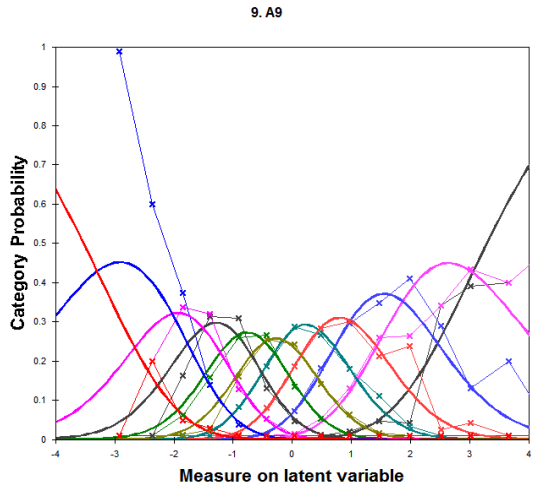

6. A6
Score on Item
Measure on latent variable

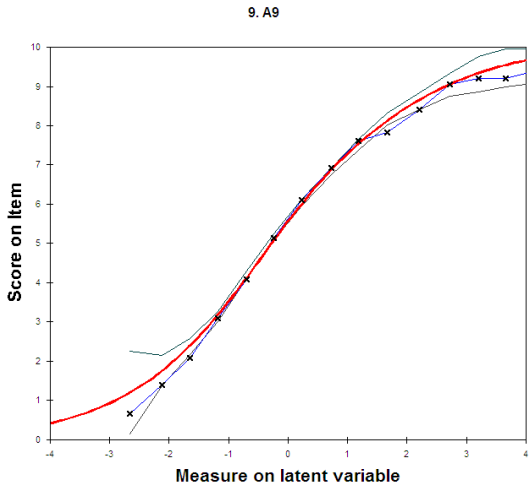

7. A7
Category Probability
Measure on latent variable

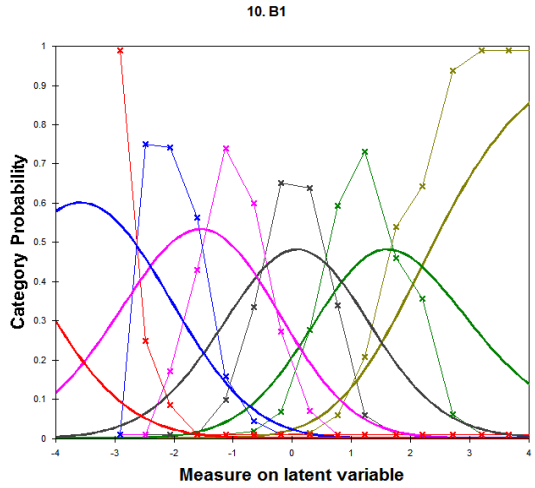

7. A7
Score on Item
Measure on latent variable

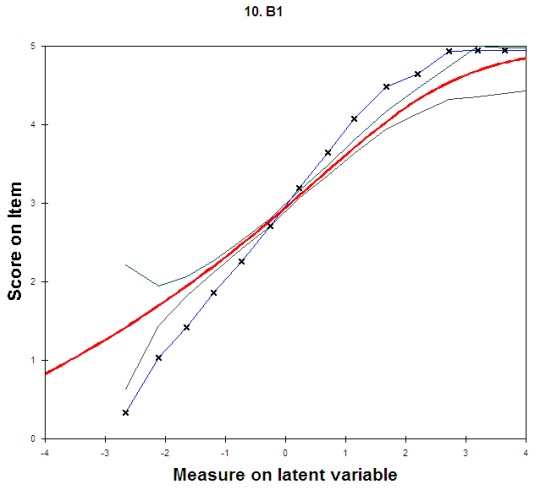

8. A8
Category Probability
Measure on latent variable

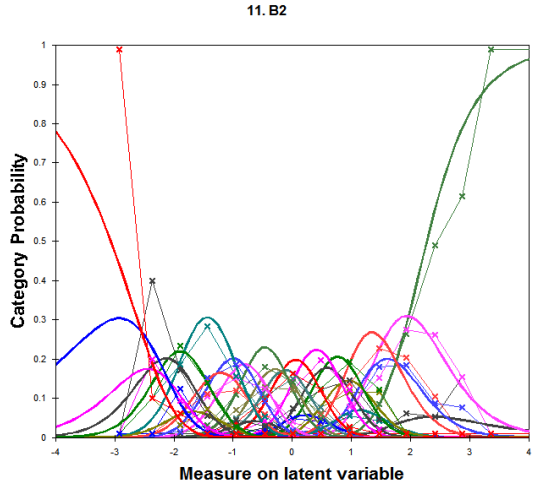

8. A8
Score on Item
Measure on latent variable

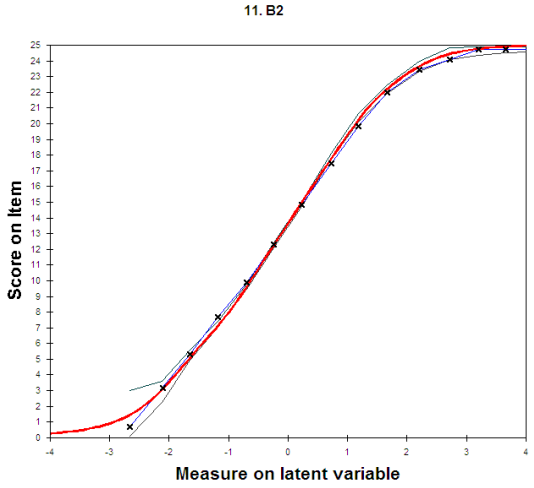

9. A9
Category Probability
Measure on latent variable

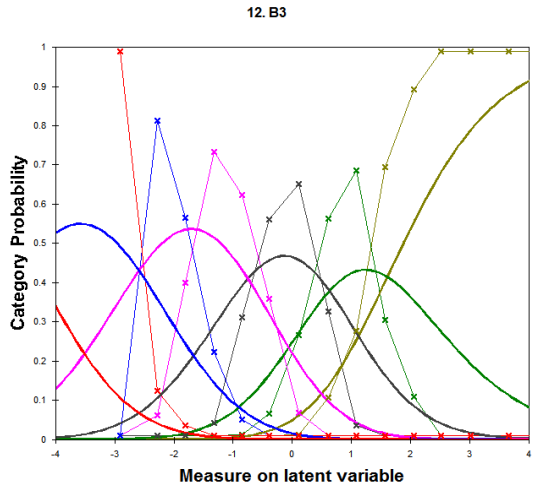

9. A9
Score on Item
Measure on latent variable

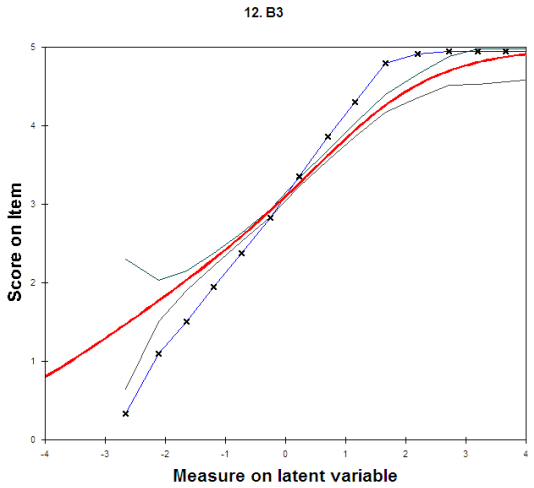

10. B1
Category Probability
Measure on latent variable

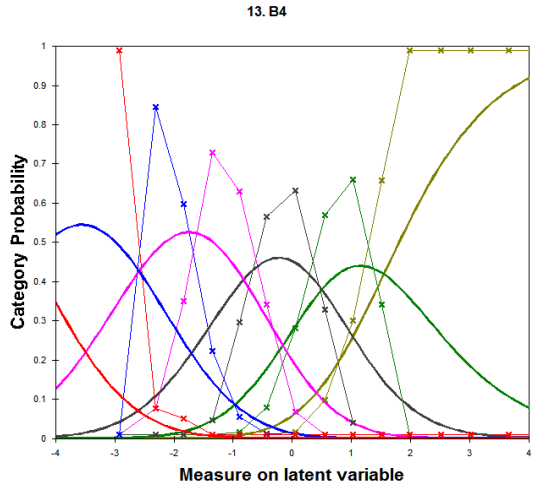

10. B1
Score on Item
Measure on latent variable

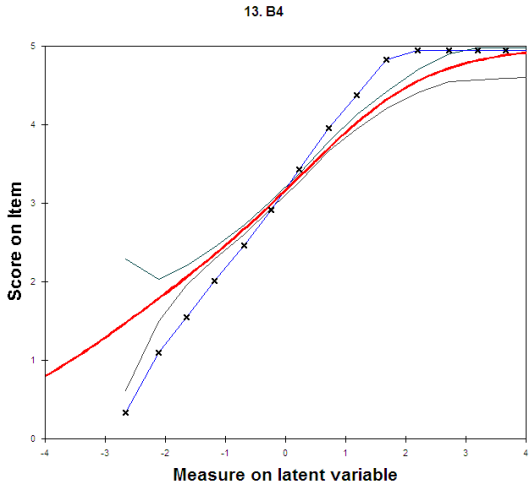

11. B2
Category Probability
Measure on latent variable


11. B2
Score on Item
Measure on latent variable


12. B3
Category Probability
Measure on latent variable

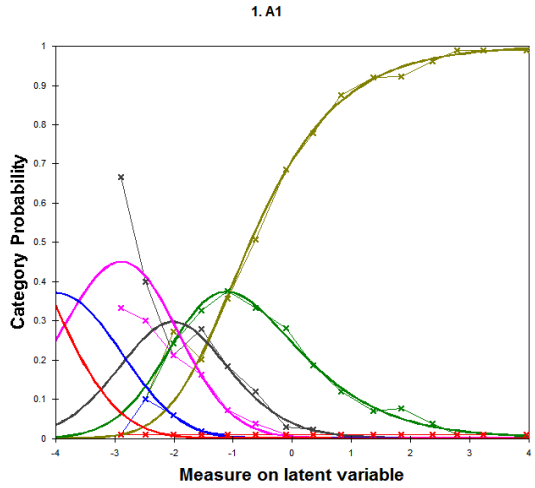

12. B3
Score on Item
Measure on latent variable

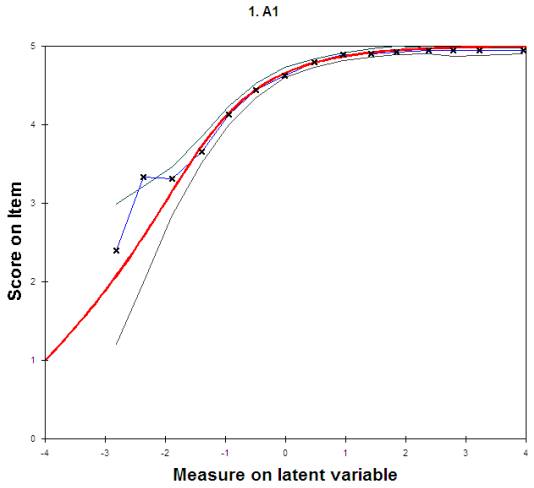

13. B4
Category Probability
Measure on latent variable

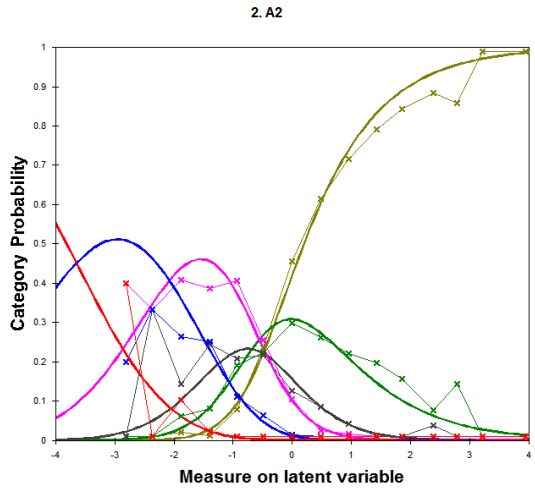

13. B4
Score on Item
Measure on latent variable

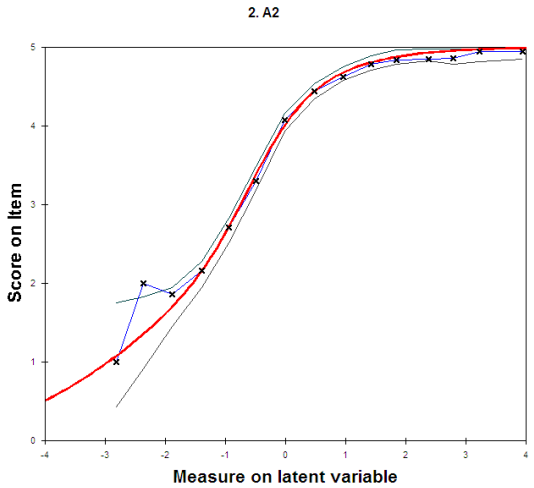
Appendix G

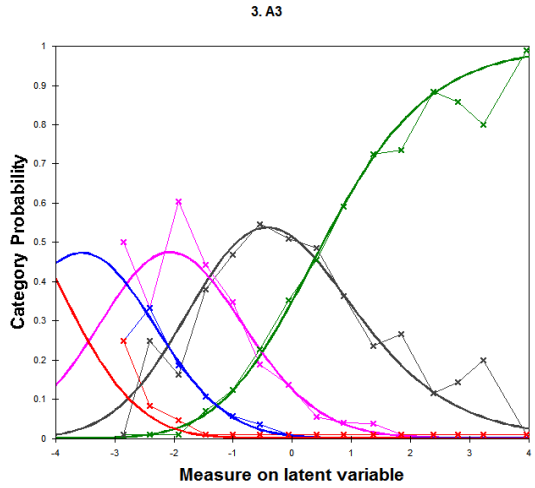
Item-level fit characteristics for AQA's GERM3 in June 2013

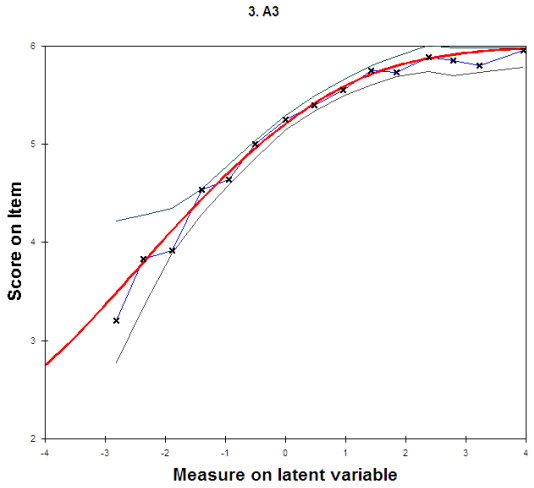


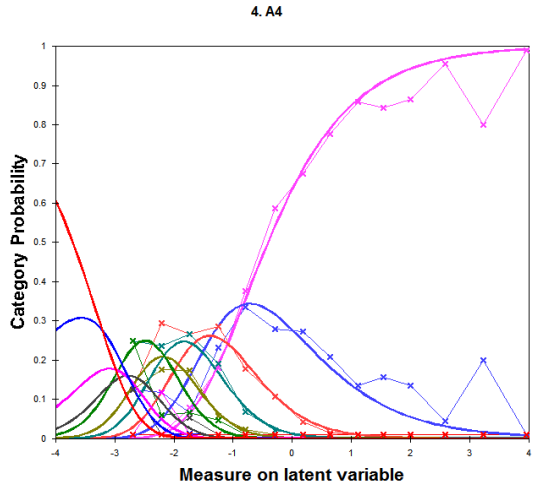


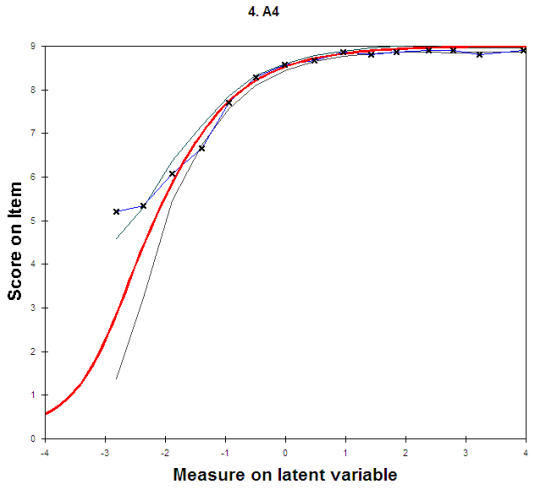


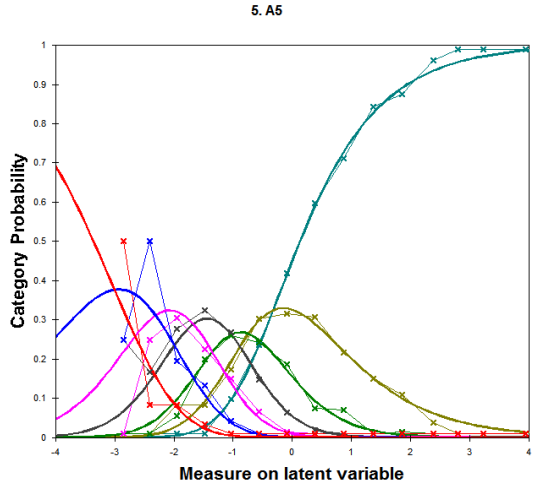


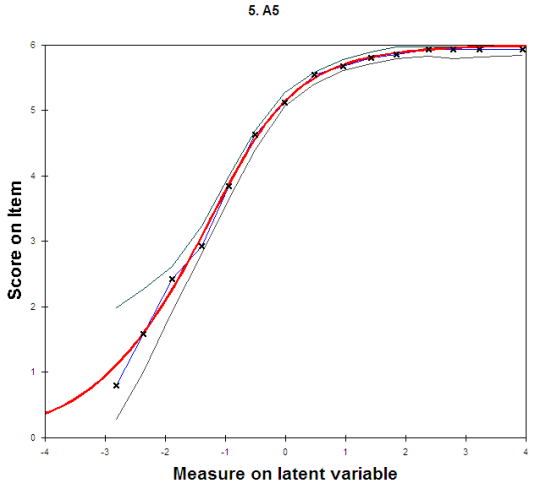


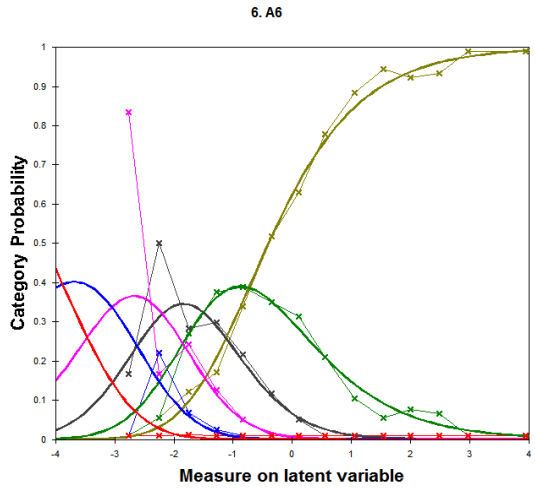


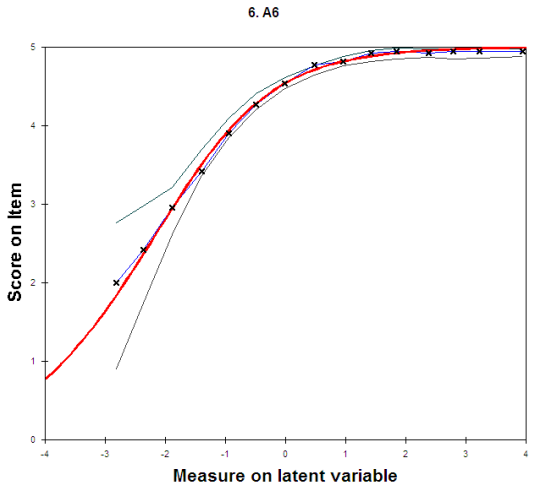

4. A4
Category Probability
Measure on latent variable

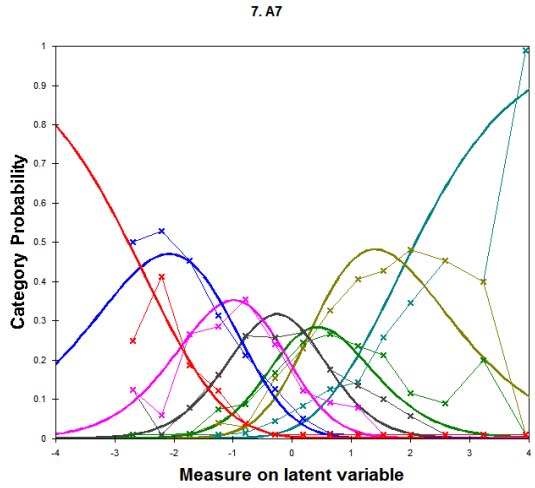

4. A4
Score on Item
Measure on latent variable

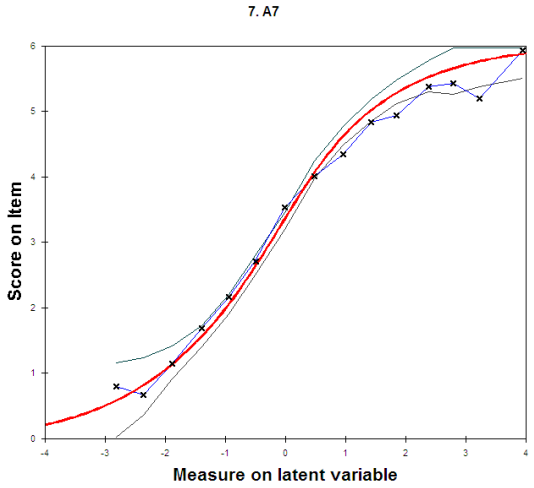

5. A5
Category Probability
Measure on latent variable

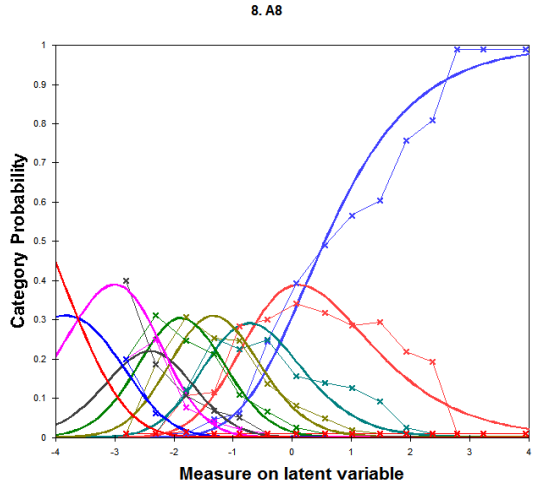

5. A5
Score on Item
Measure on latent variable

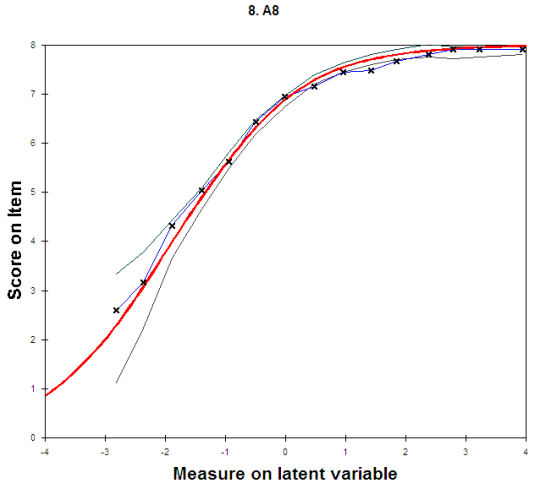

6. A6
Category Probability
Measure on latent variable

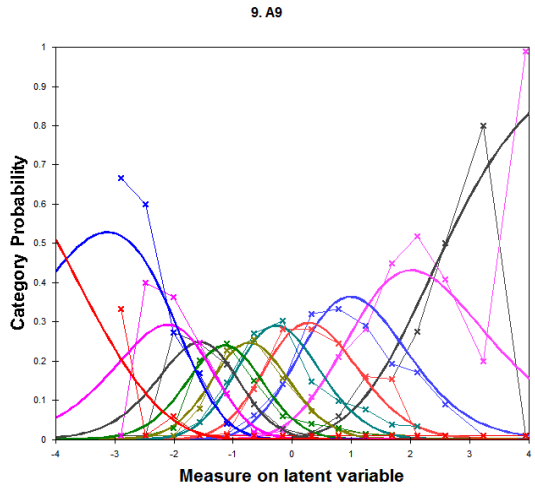

6. A6
Score on Item
Measure on latent variable

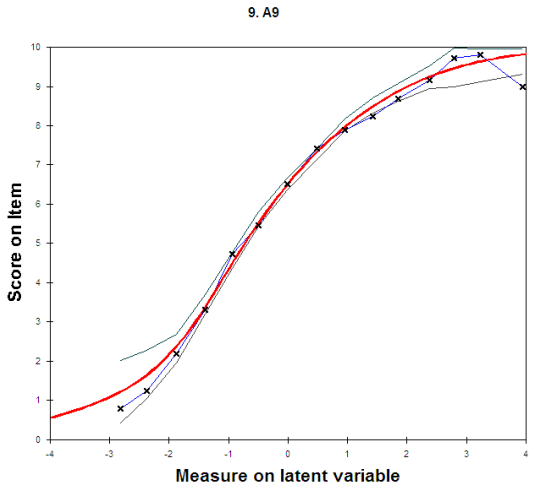

7. A7
Category Probability
Measure on latent variable

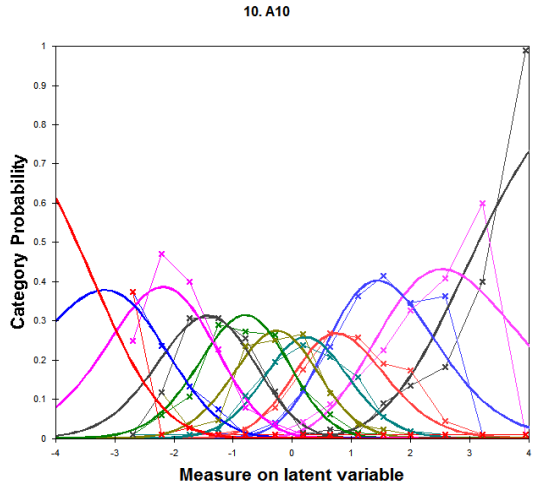

7. A7
Score on Item
Measure on latent variable

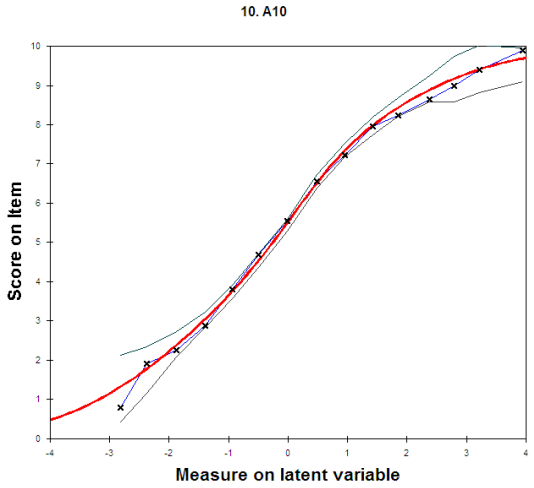

8. A8
Category Probability
Measure on latent variable

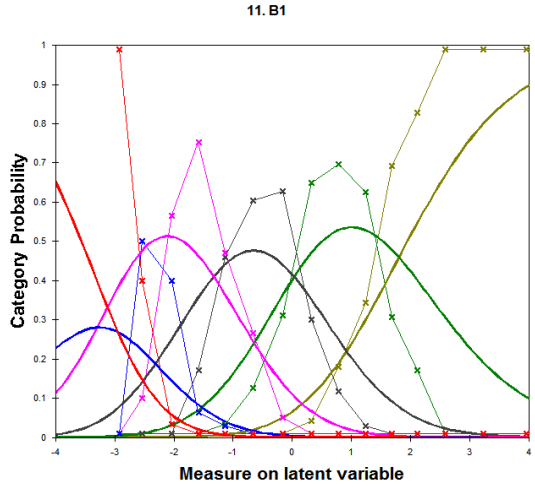

8. A8
Score on Item
Measure on latent variable

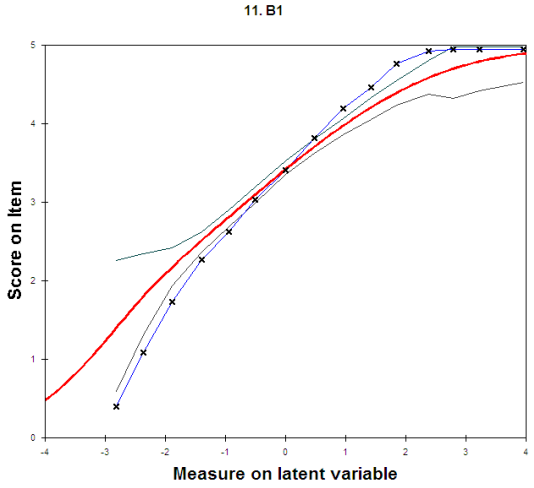

9. A9
Category Probability
Measure on latent variable

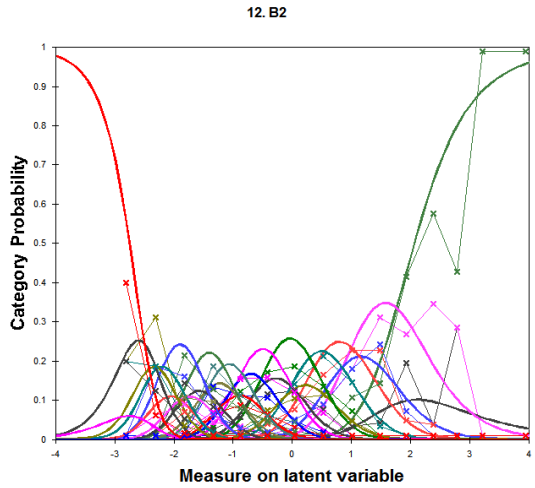

9. A9
Score on Item
Measure on latent variable

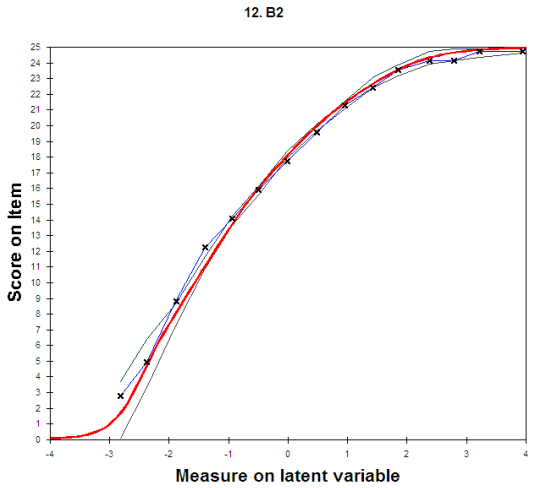

10. A10
Category Probability
Measure on latent variable

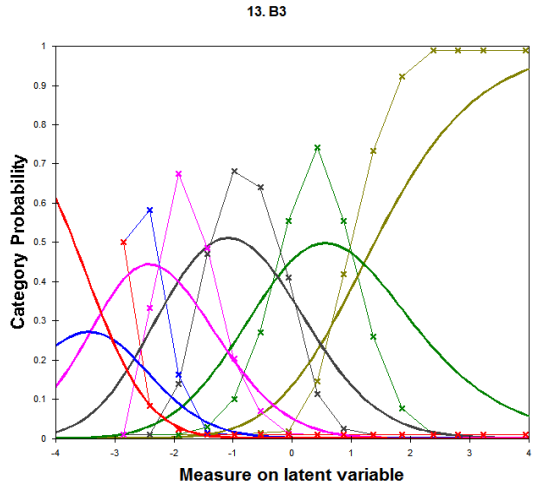

10. A10
Score on Item
Measure on latent variable

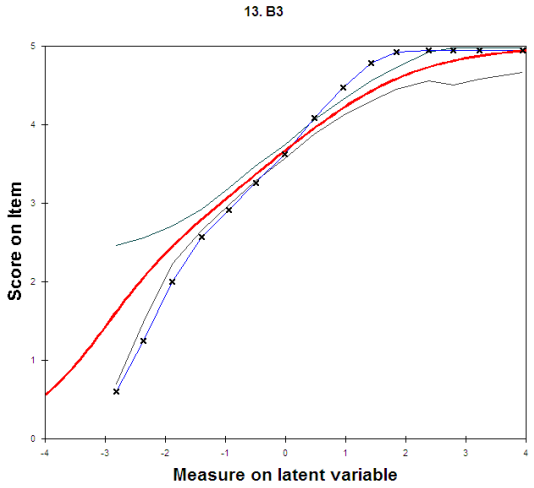

11. B1
Category Probability
Measure on latent variable

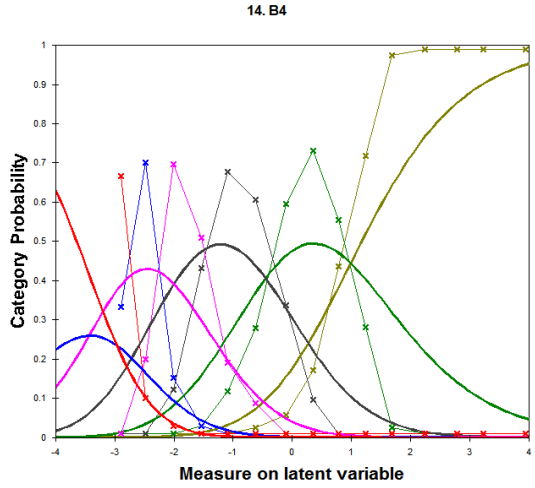

11. B1
Score on Item
Measure on latent variable

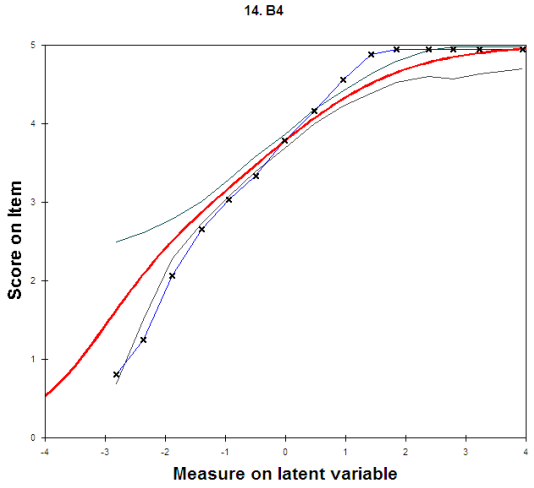

12. B2
Category Probability
Measure on latent variable


12. B2
Score on Item
Measure on latent variable


13. B3
Category Probability
Measure on latent variable

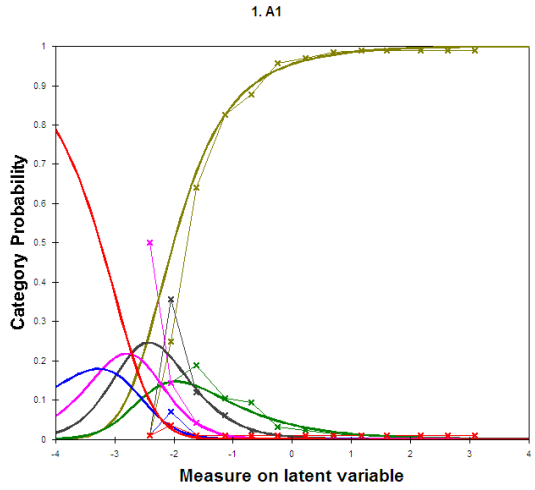

13. B3
Score on Item
Measure on latent variable

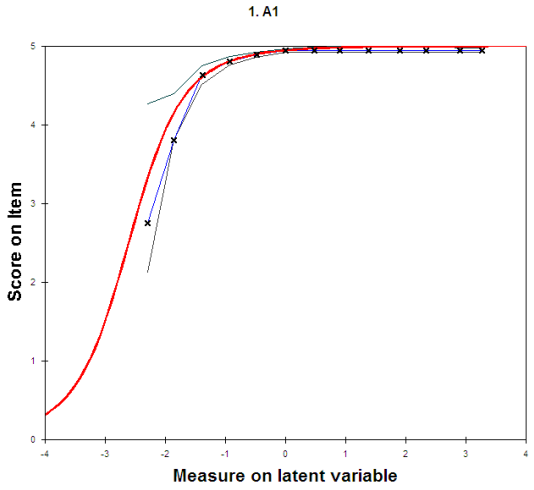

14. B4
Category Probability
Measure on latent variable

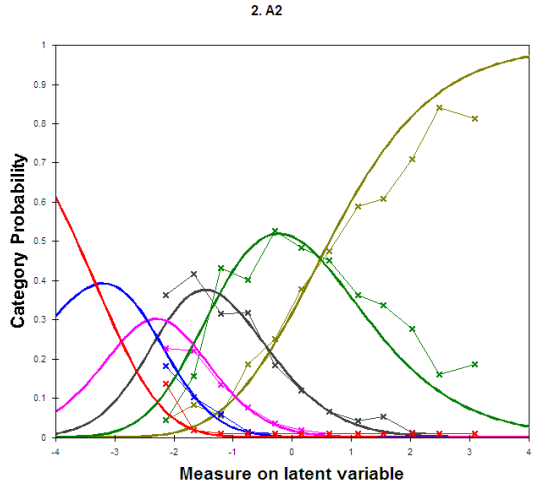

14. B4
Score on Item
Measure on latent variable

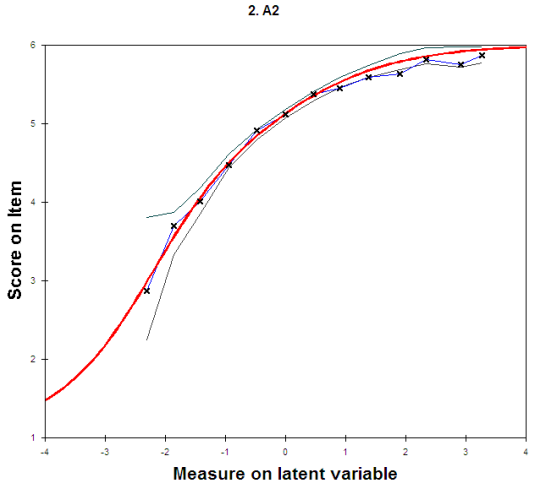
Appendix H

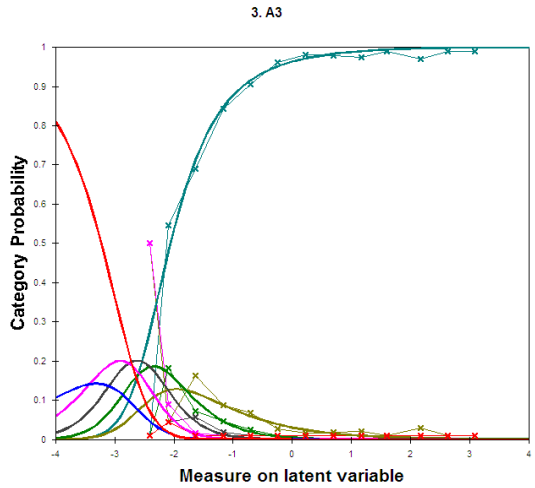
Item-level fit characteristics for AQA's SPAN3 in June 2013

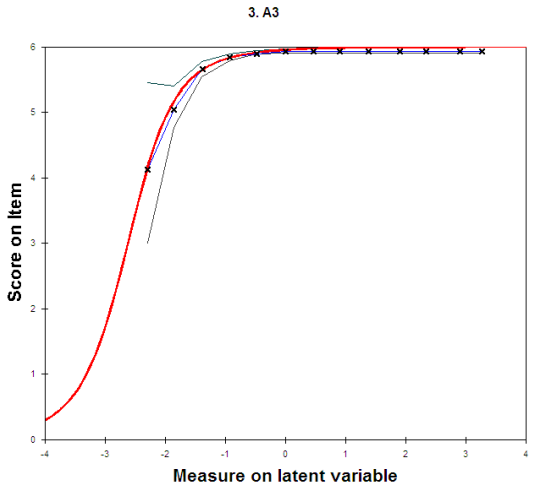


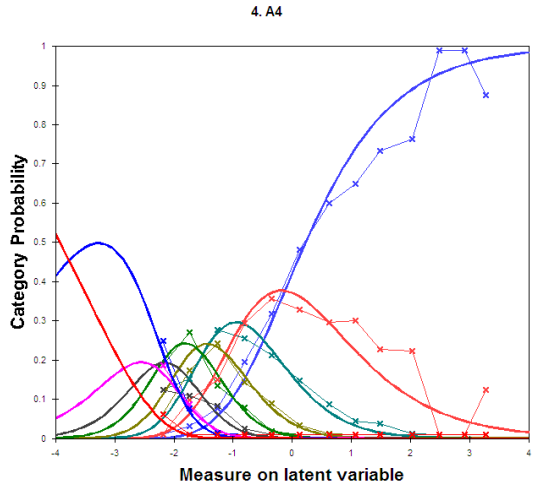


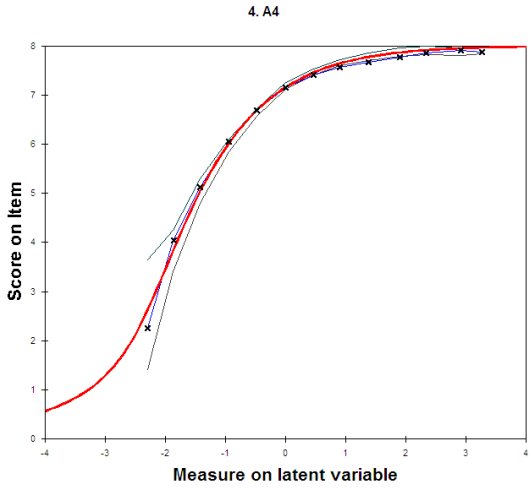


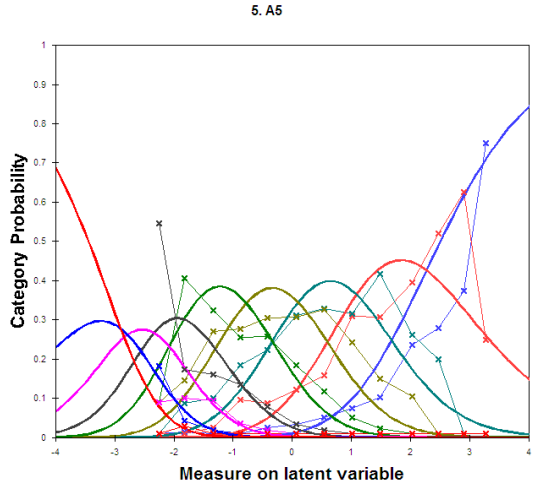


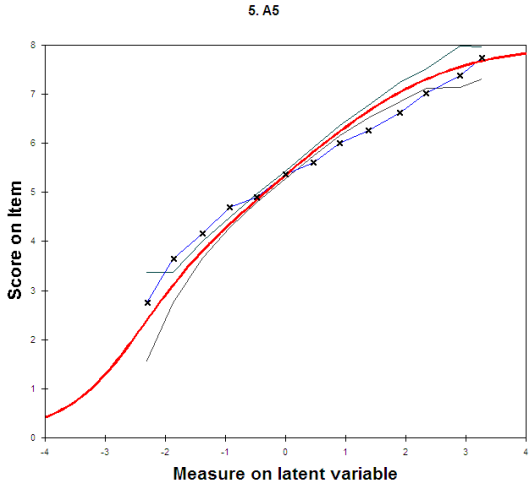


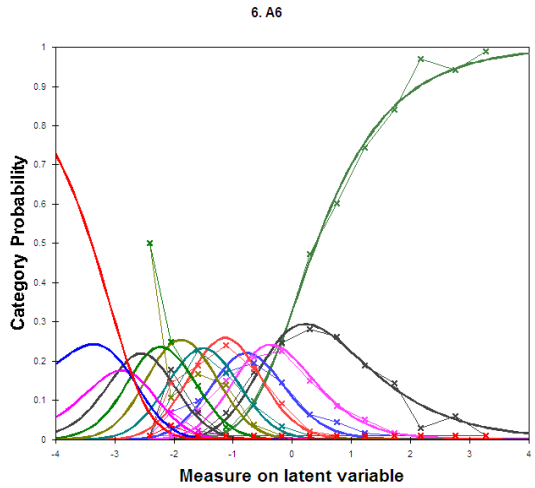


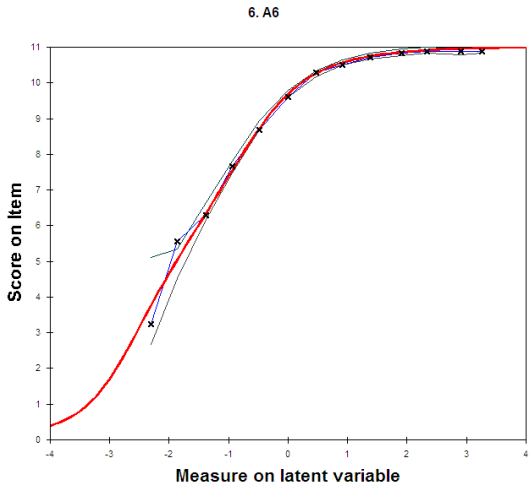

4. A4
Category Probability
Measure on latent variable

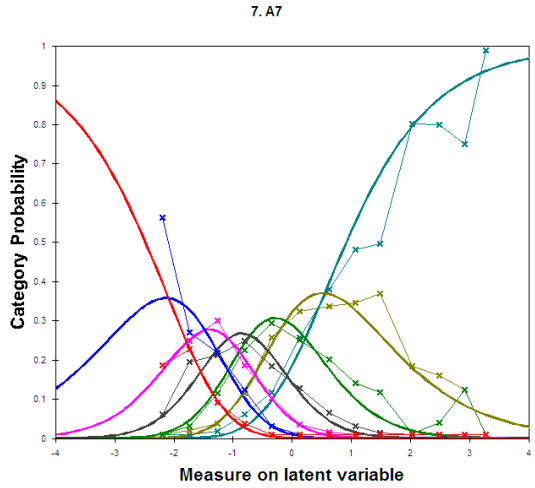

4. A4
Score on Item
Measure on latent variable

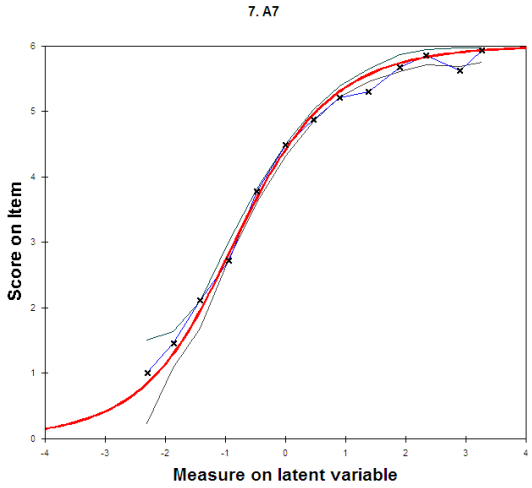

5. A5
Category Probability
Measure on latent variable

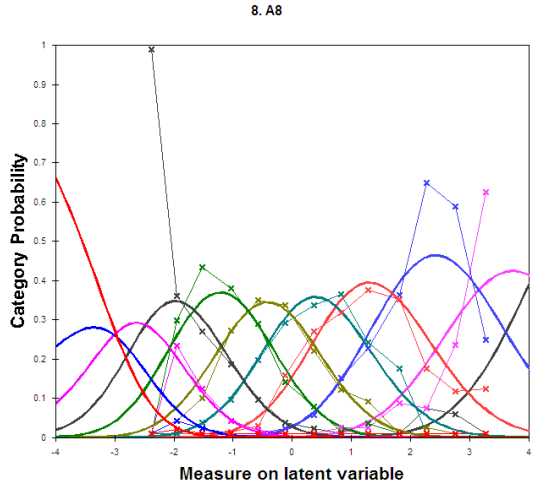

5. A5
Score on Item
Measure on latent variable

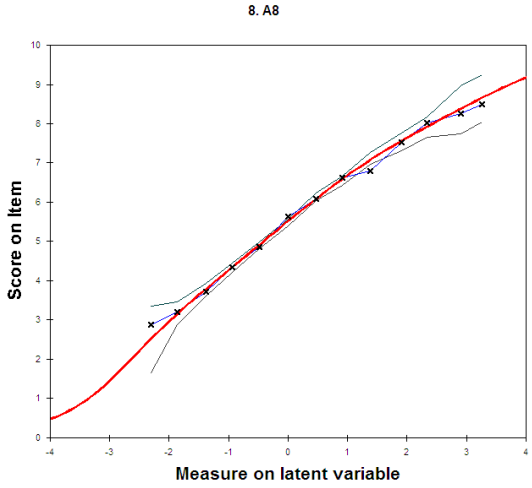

6. A6
Category Probability
Measure on latent variable

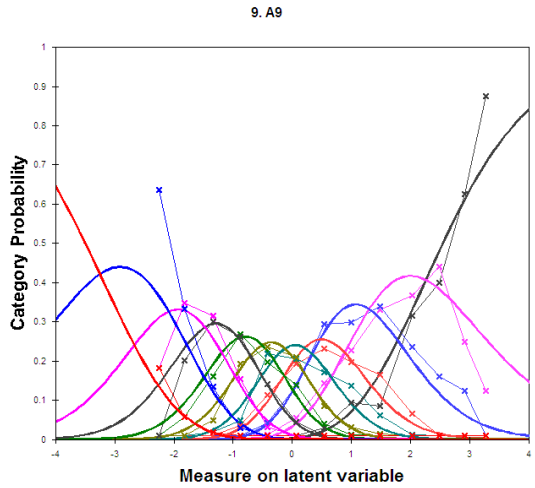

6. A6
Score on Item
Measure on latent variable

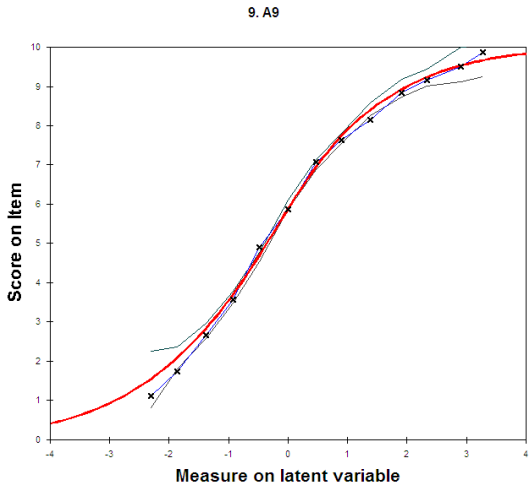

7. A7
Category Probability
Measure on latent variable

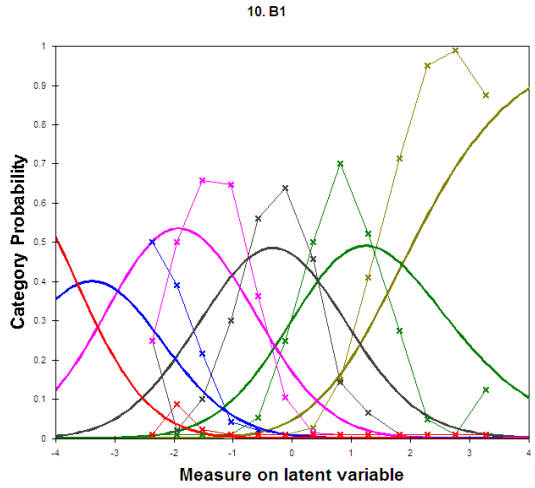

7. A7
Score on Item
Measure on latent variable

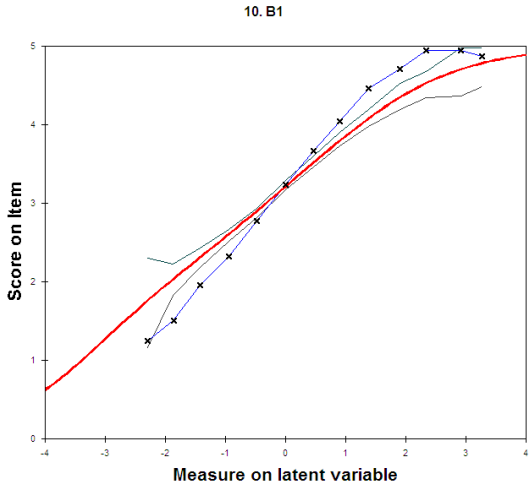

8. A8
Category Probability
Measure on latent variable

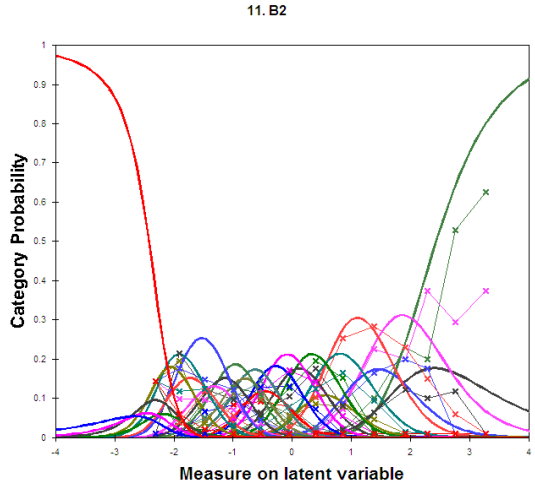

8. A8
Score on Item
Measure on latent variable

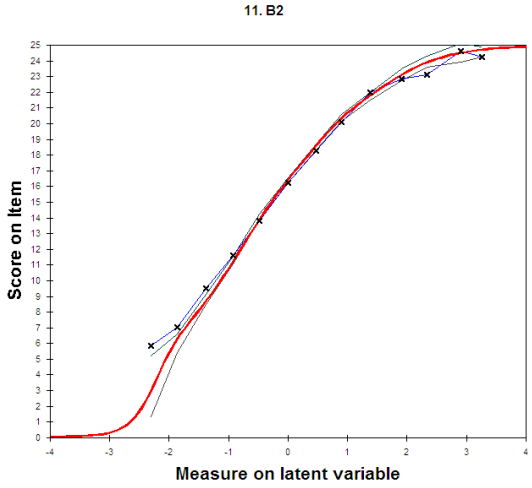

9. A9
Category Probability
Measure on latent variable

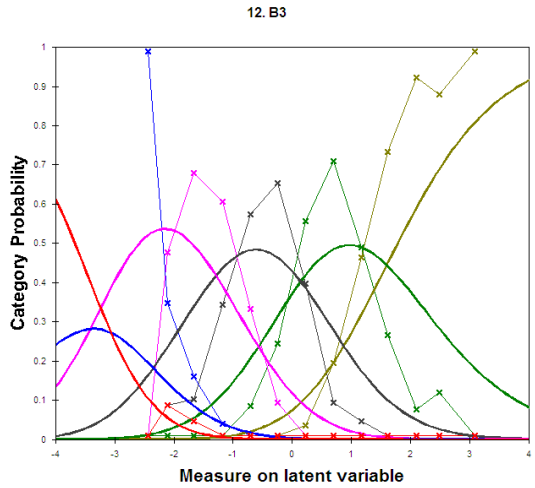

9. A9
Score on Item
Measure on latent variable

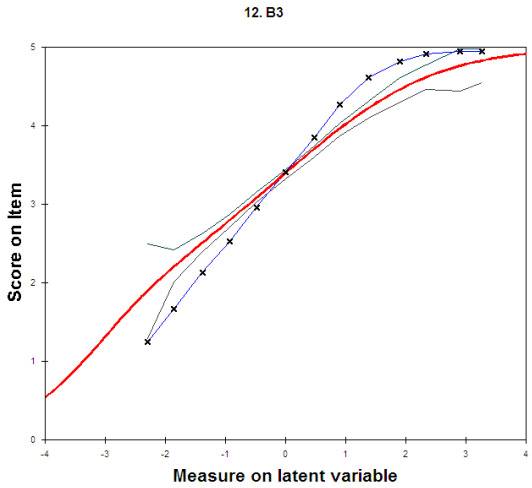

10. B1
Category Probability
Measure on latent variable

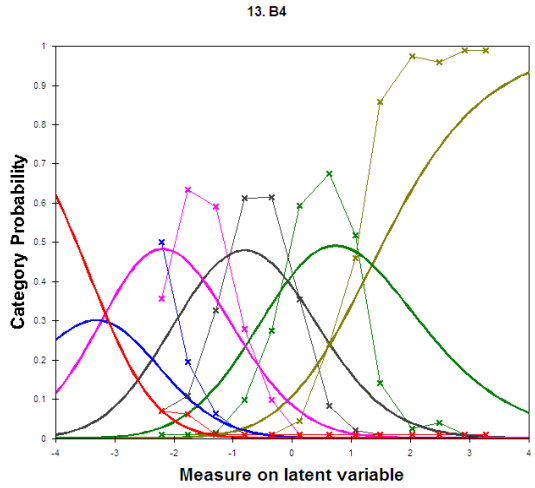

10. B1
Score on Item
Measure on latent variable

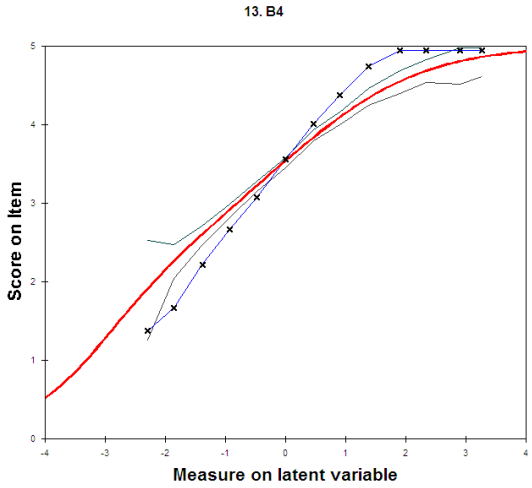

11. B2
Category Probability
Measure on latent variable


11. B2
Score on Item
Measure on latent variable


12. B3
Category Probability
Measure on latent variable

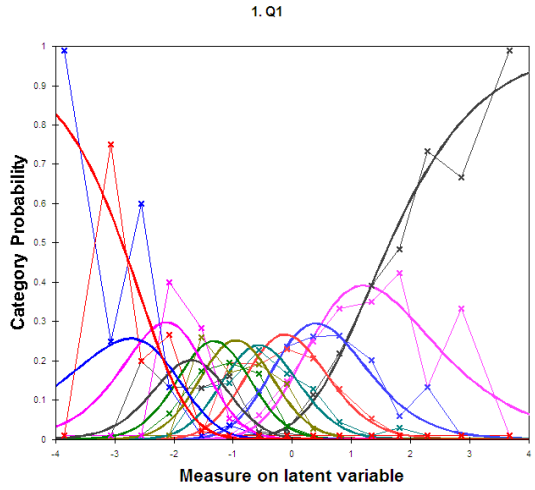

12. B3
Score on Item
Measure on latent variable

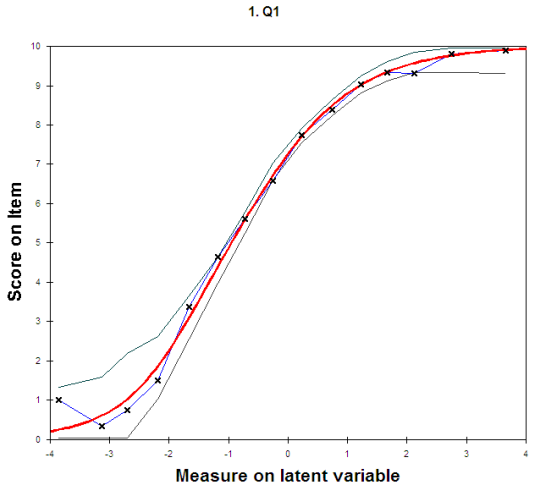

13. B4
Category Probability
Measure on latent variable

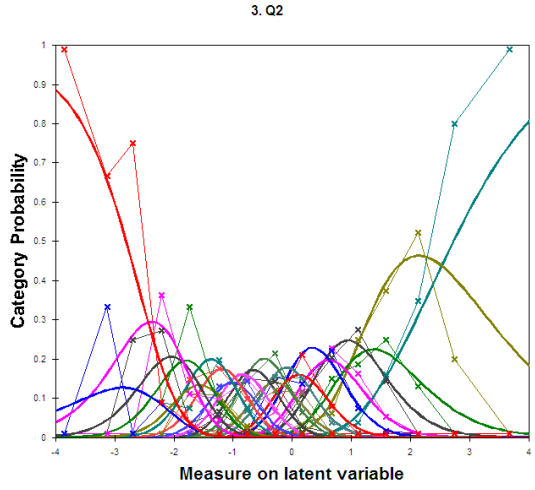

13. B4
Score on Item
Measure on latent variable

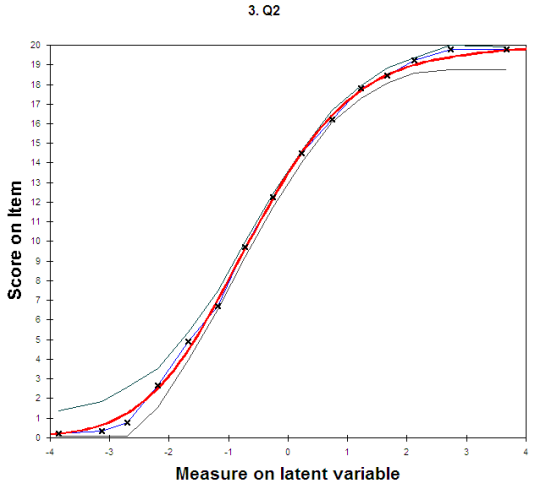
Appendix I

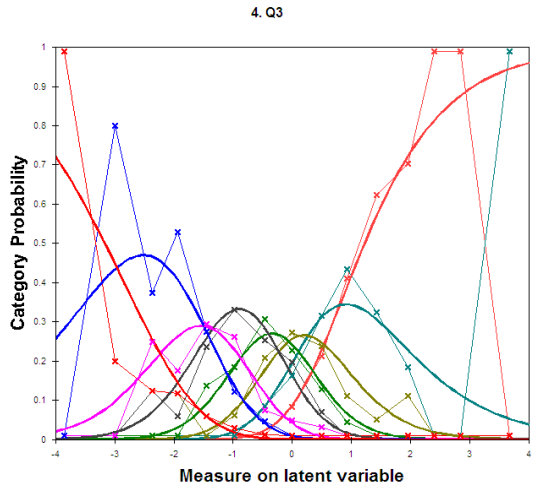
Item-level fit characteristics for OCR's F704 in June 2013

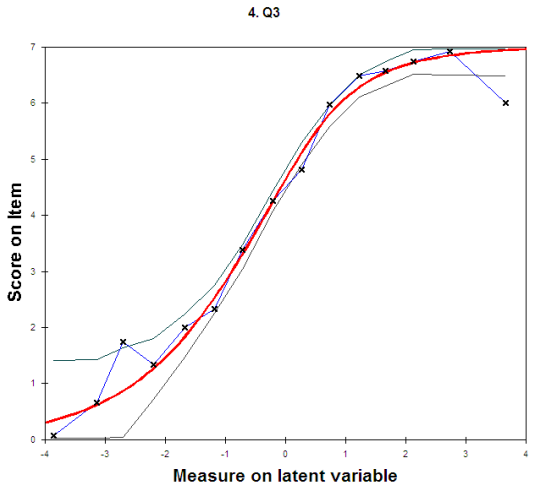


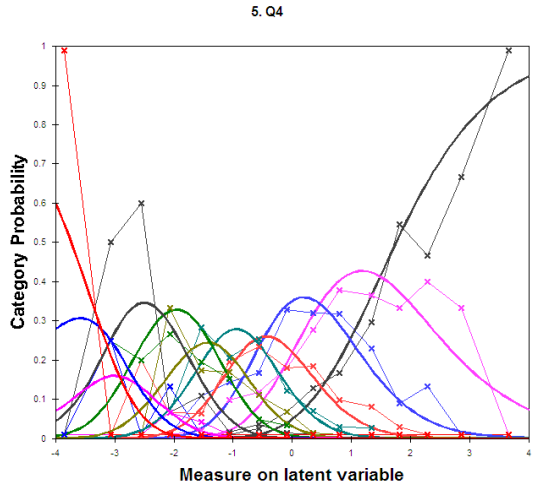


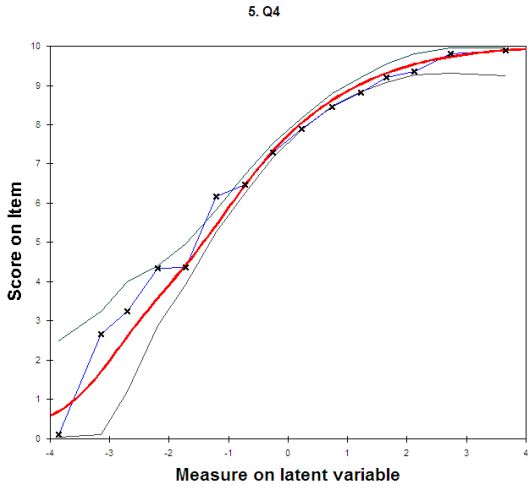


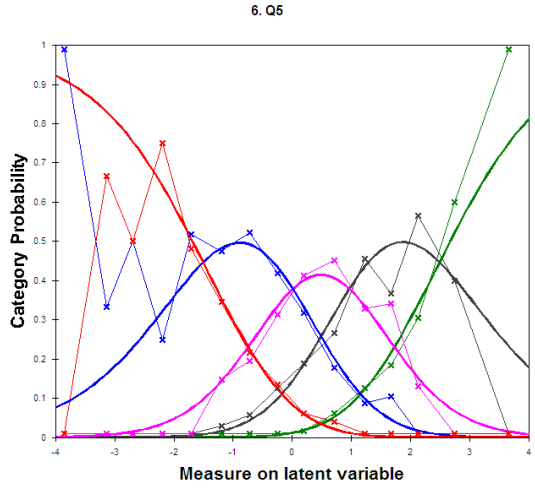


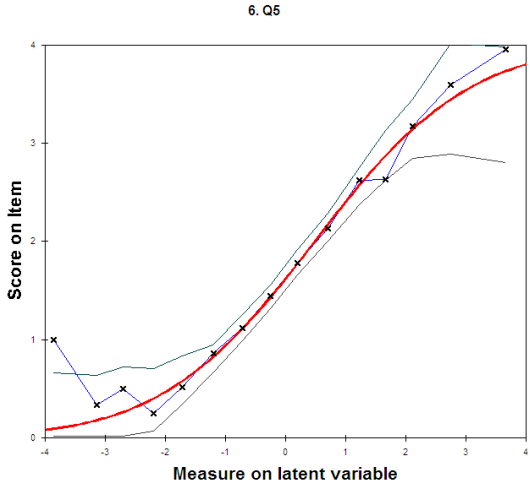


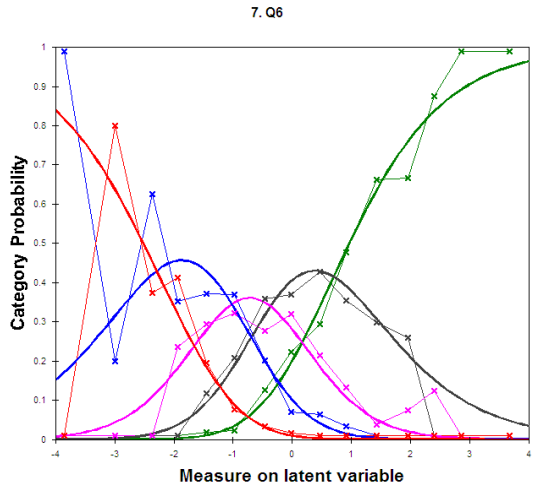


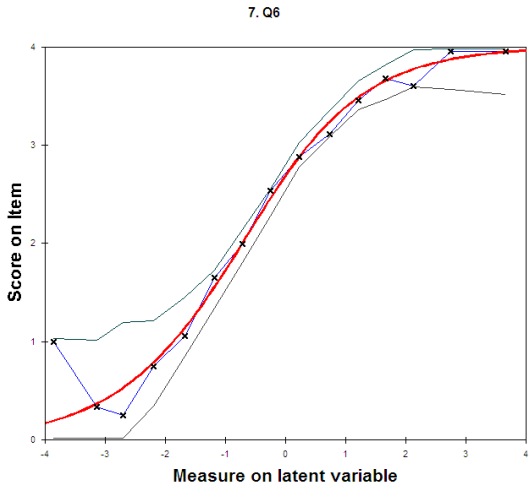

5. Q4
Category Probability
Measure on latent variable

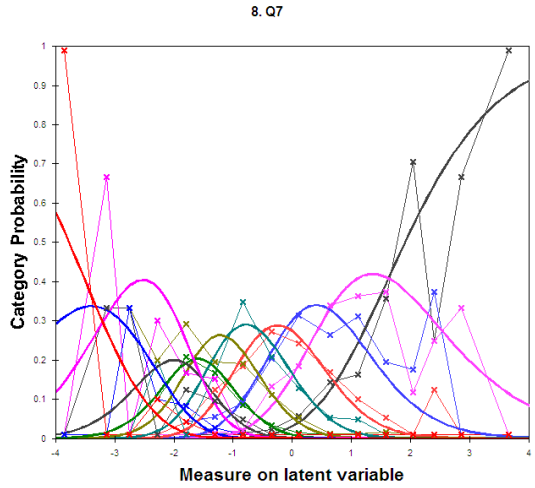

5. Q4
Score on Item
Measure on latent variable

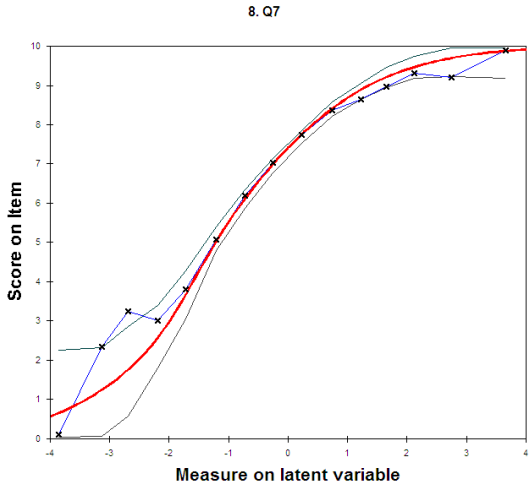

6. Q5
Category Probability
Measure on latent variable

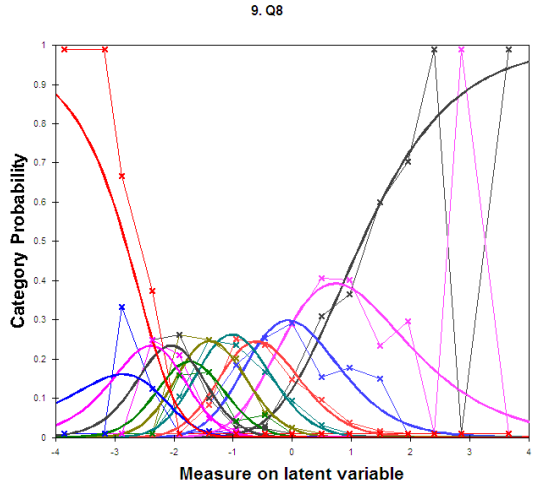

6. Q5
Score on Item
Measure on latent variable

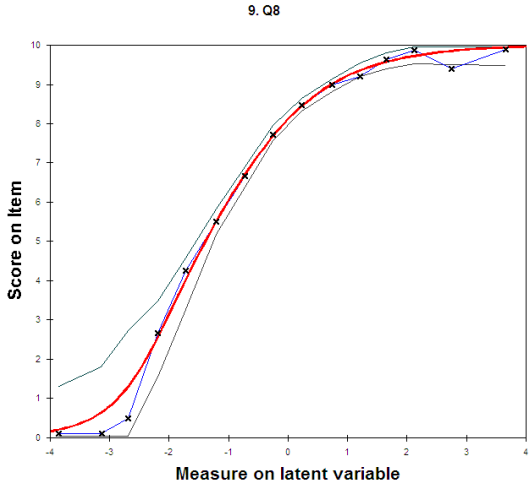

7. Q6
Category Probability
Measure on latent variable

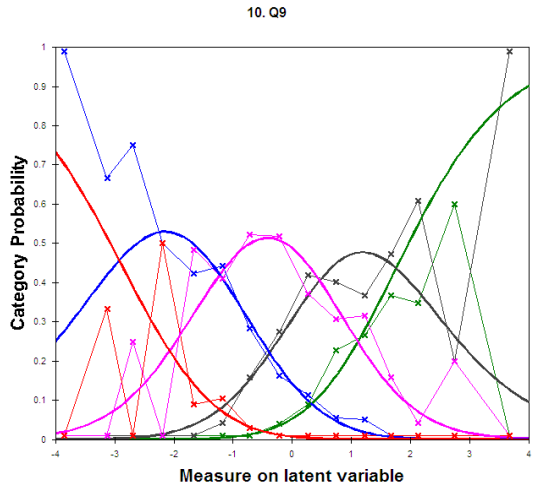

7. Q6
Score on Item
Measure on latent variable

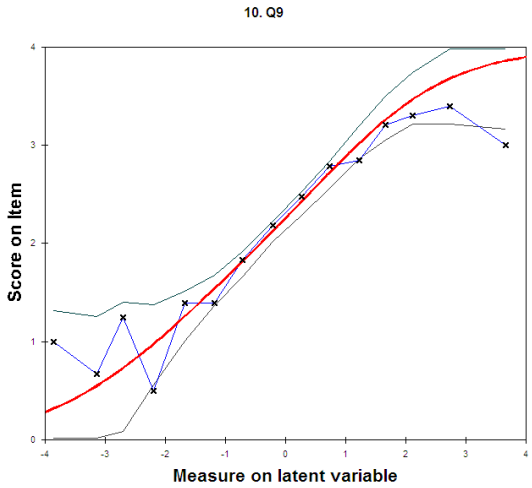

8. Q7
Category Probability
Measure on latent variable

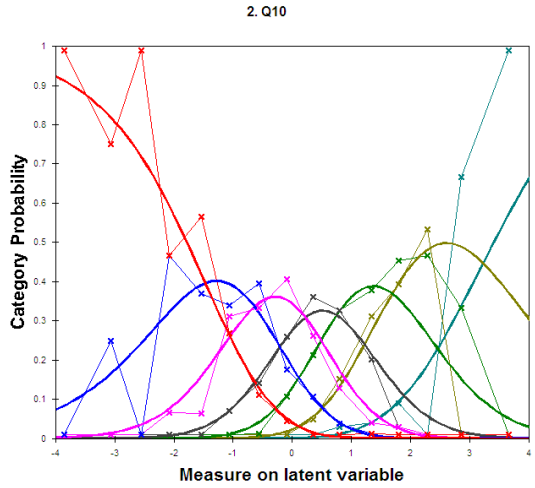

8. Q7
Score on Item
Measure on latent variable

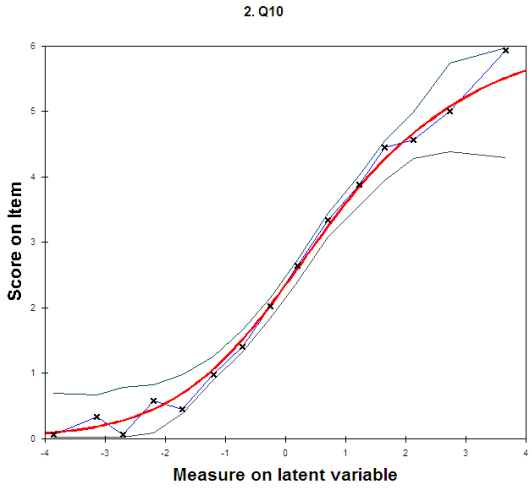

9. Q8
Category Probability
Measure on latent variable

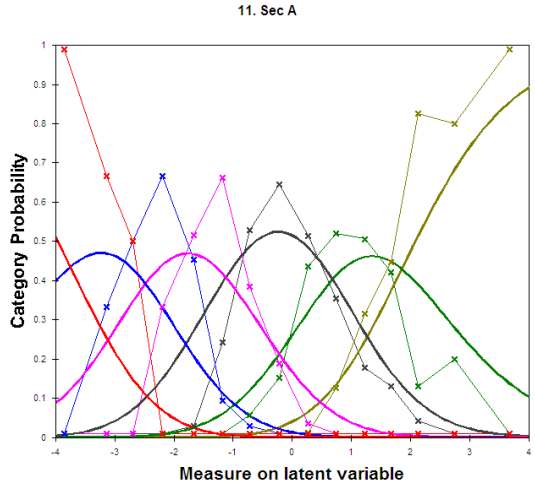

9. Q8
Score on Item
Measure on latent variable

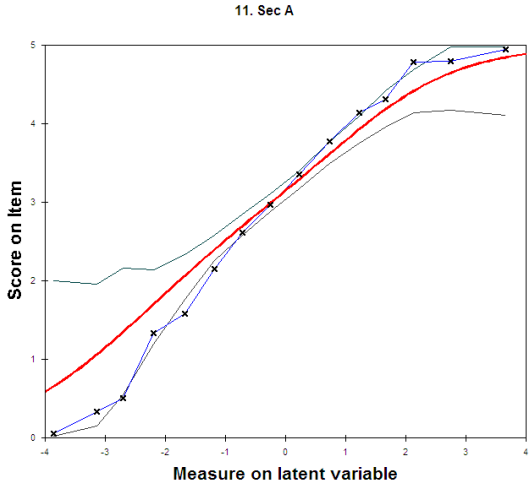

10. Q9
Category Probability
Measure on latent variable

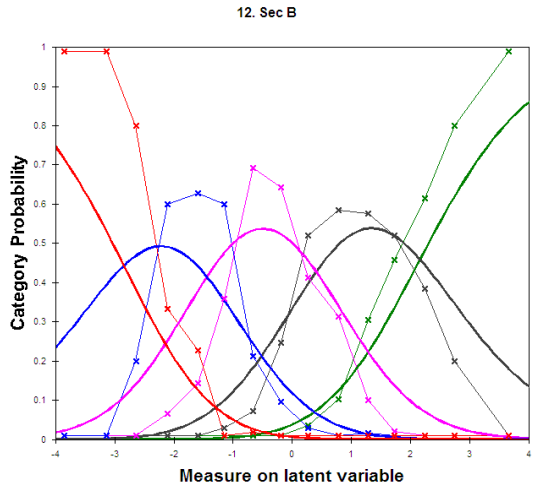

10. Q9
Score on Item
Measure on latent variable

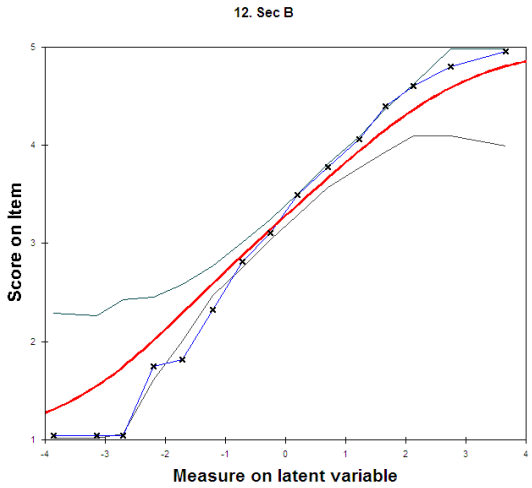

2. Q10
Category Probability
Measure on latent variable


2. Q10
Score on Item
Measure on latent variable


11. Sec A
Category Probability
Measure on latent variable

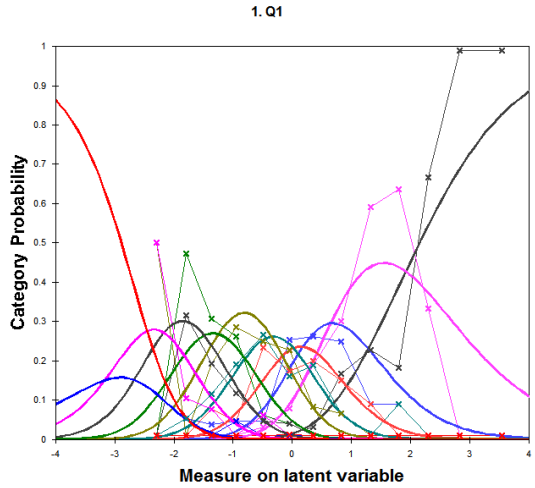

11. Sec A
Score on Item
Measure on latent variable

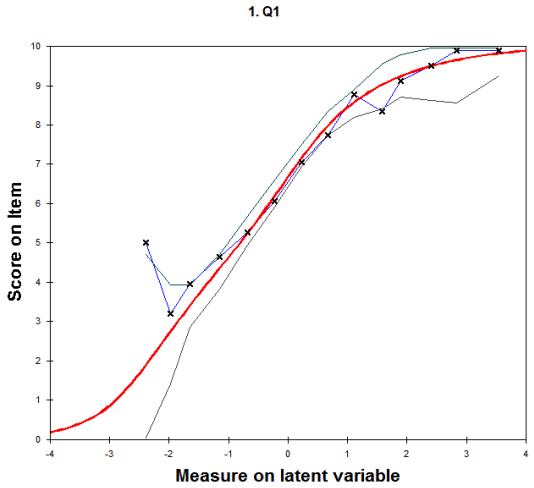

12. Sec B
Category Probability
Measure on latent variable

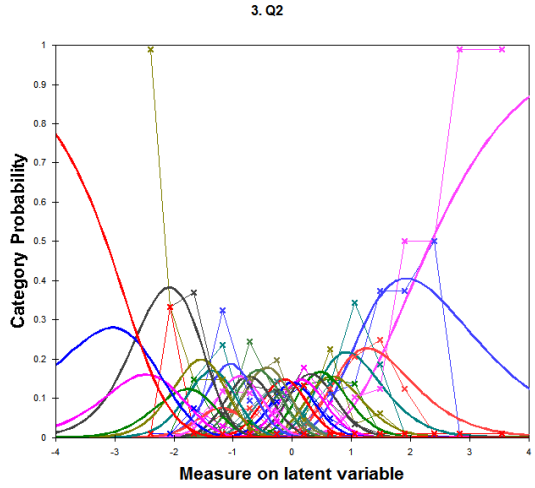

12. Sec B
Score on Item
Measure on latent variable

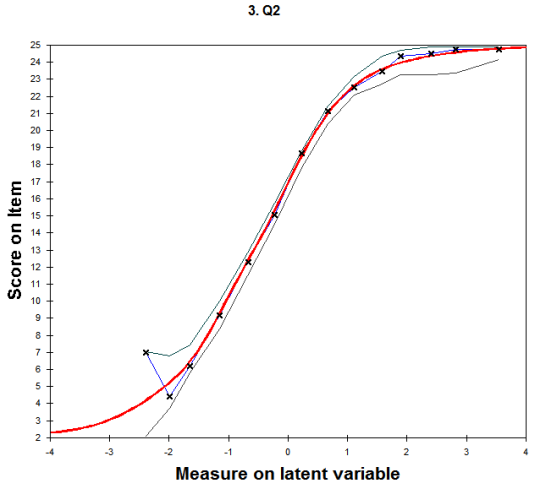
Appendix J

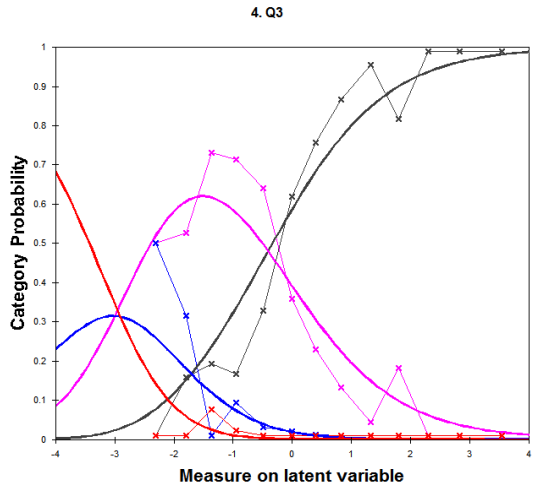
Item-level fit characteristics for OCR's F714 in June 2013

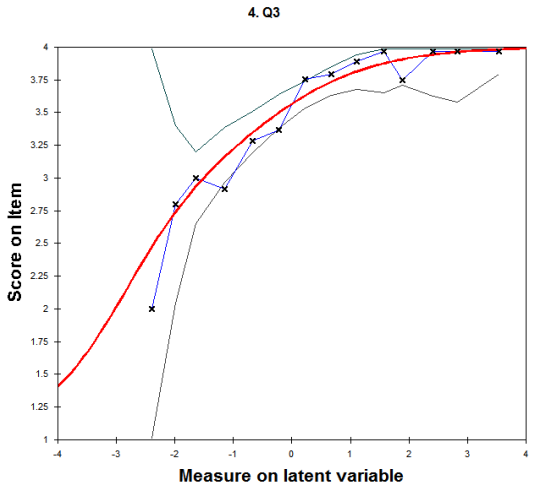


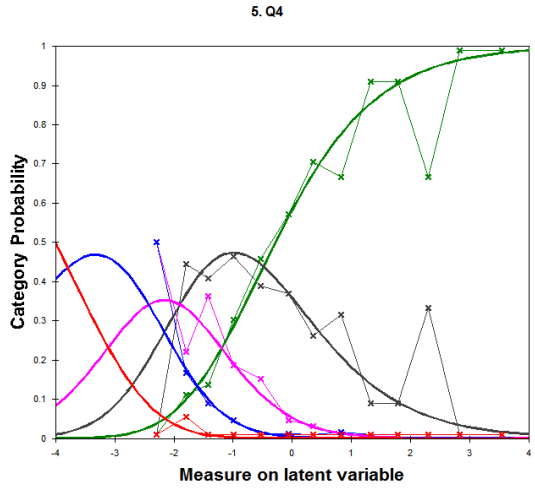


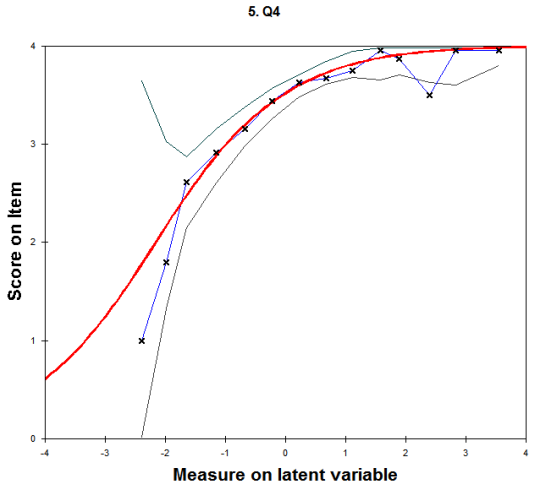


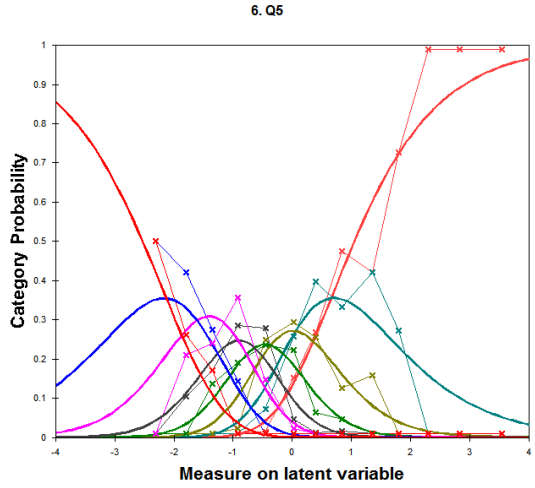


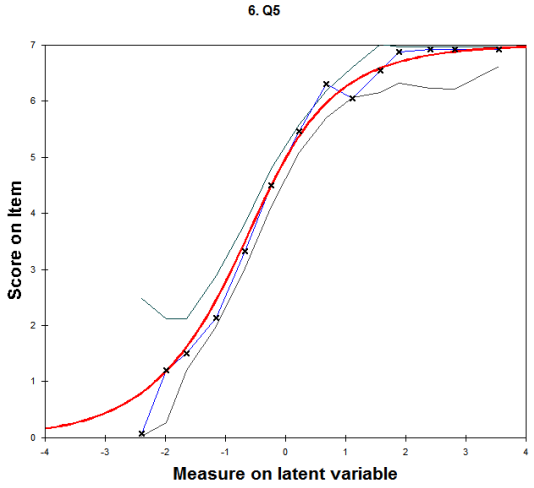


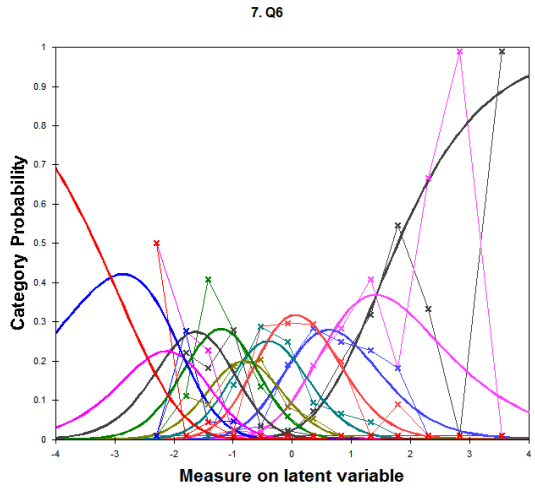


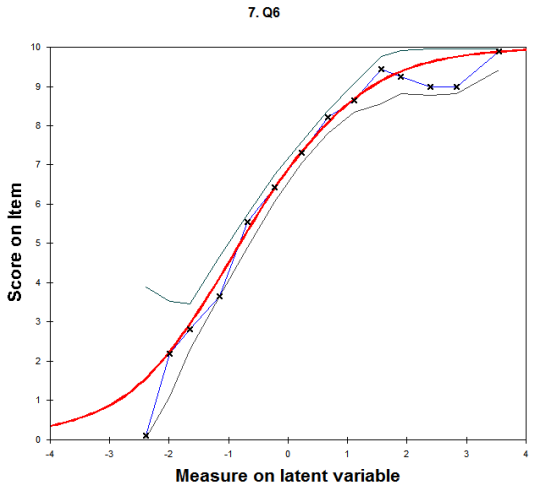

5. Q4
Category Probability
Measure on latent variable

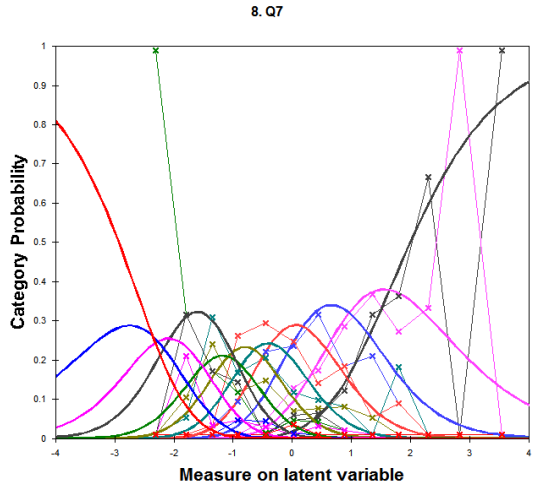

5. Q4
Score on Item
Measure on latent variable

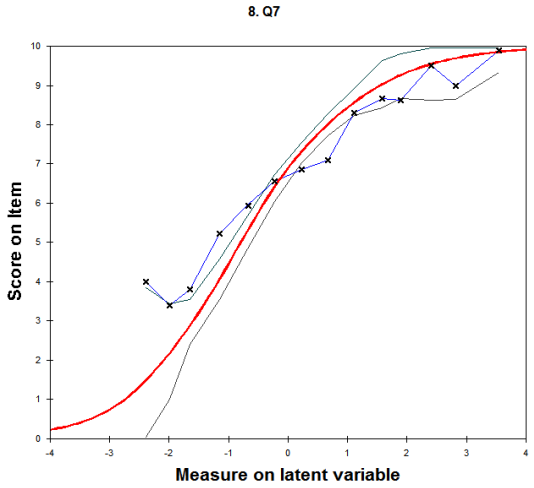

6. Q5
Category Probability
Measure on latent variable

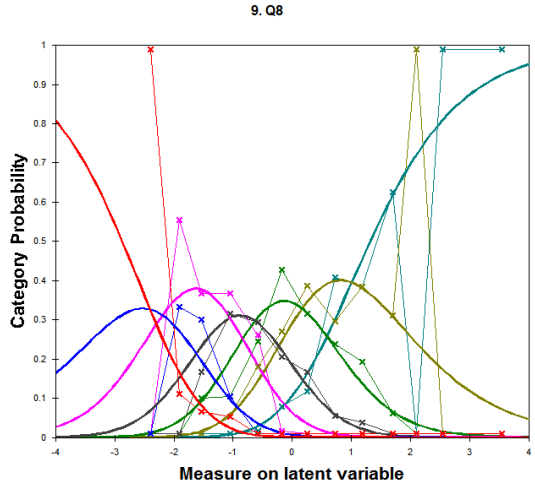

6. Q5
Score on Item
Measure on latent variable

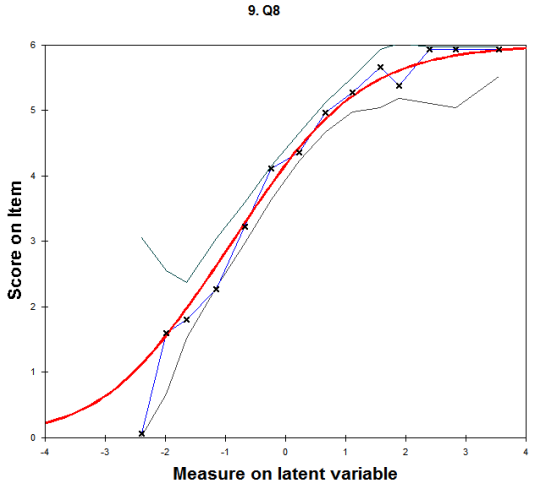

7. Q6
Category Probability
Measure on latent variable

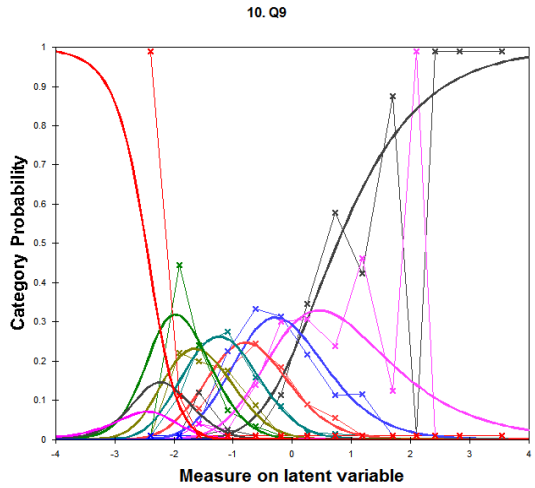

7. Q6
Score on Item
Measure on latent variable

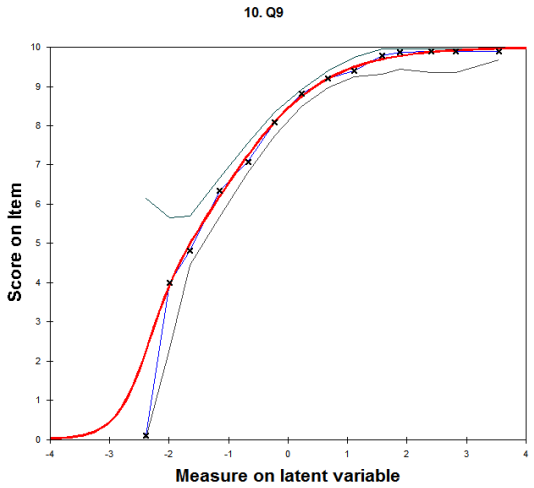

8. Q7
Category Probability
Measure on latent variable

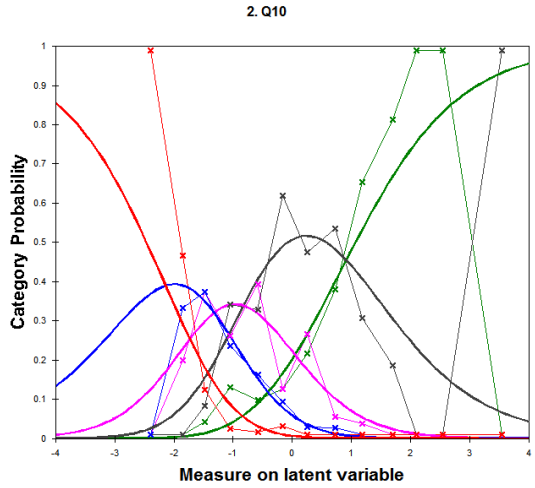

8. Q7
Score on Item
Measure on latent variable

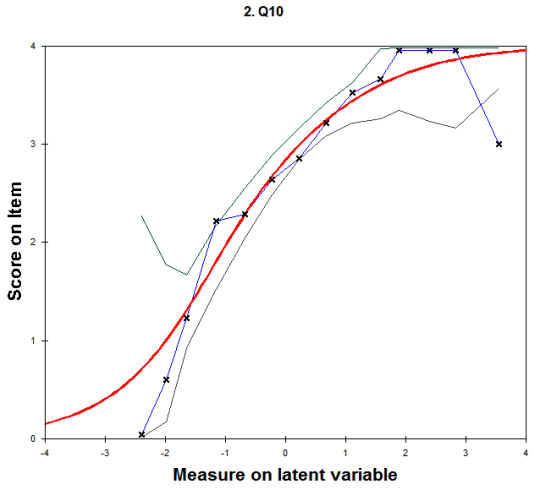

9. Q8
Category Probability
Measure on latent variable

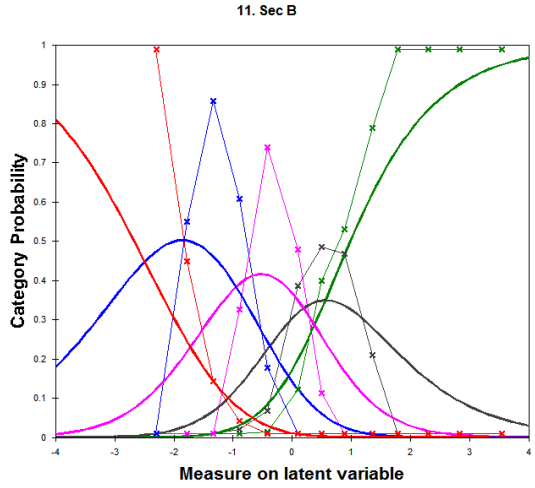

9. Q8
Score on Item
Measure on latent variable

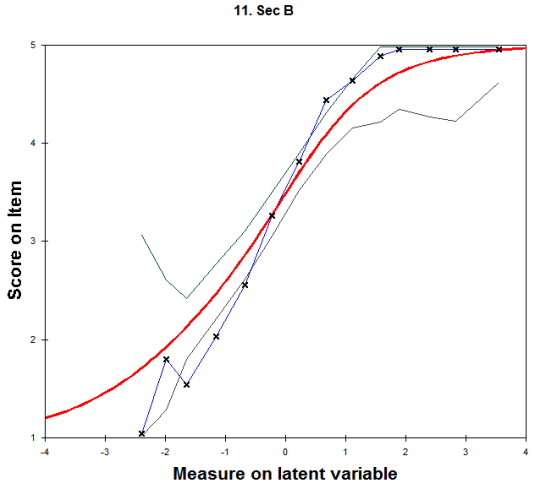

10. Q9
Category Probability
Measure on latent variable


10. Q9
Score on Item
Measure on latent variable


2. Q10
Category Probability
Measure on latent variable

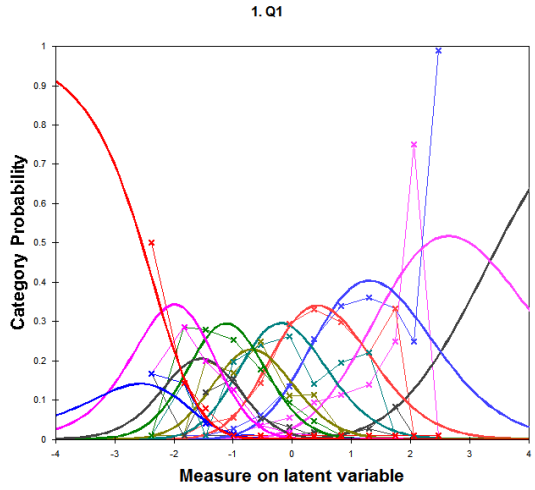

2. Q10
Score on Item
Measure on latent variable

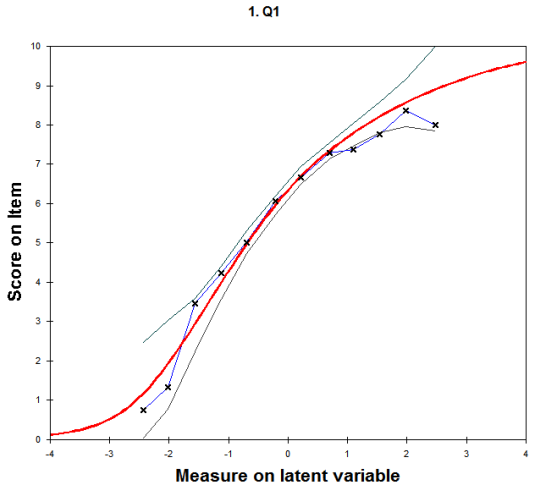

11. Sec B
Category Probability
Measure on latent variable

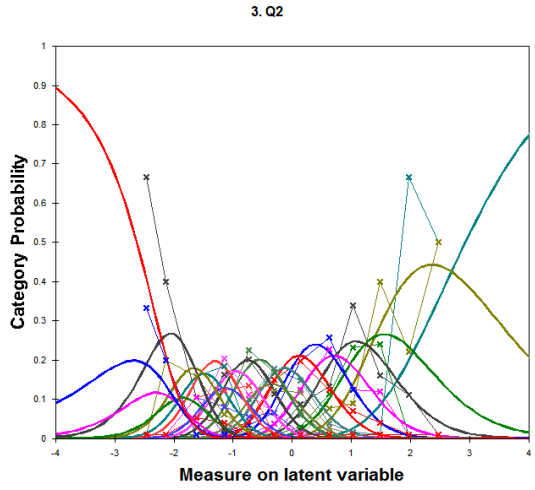

11. Sec B
Score on Item
Measure on latent variable

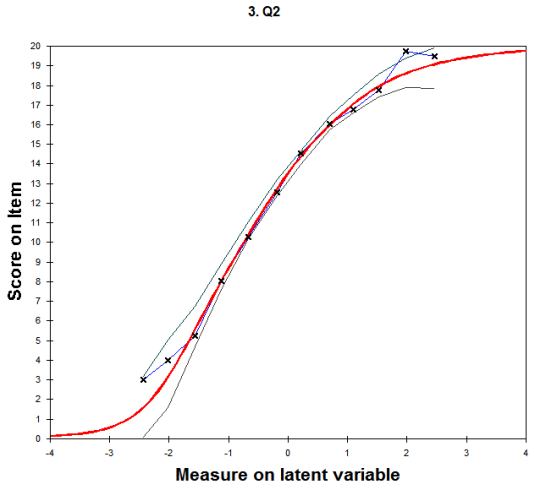
Appendix K

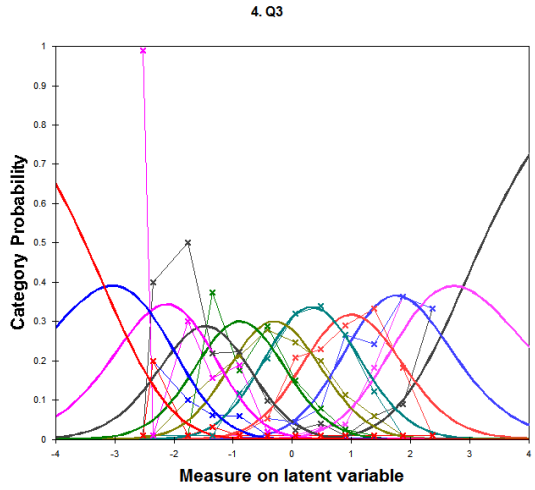
Item-level fit characteristics for OCR's F724 in June 2013

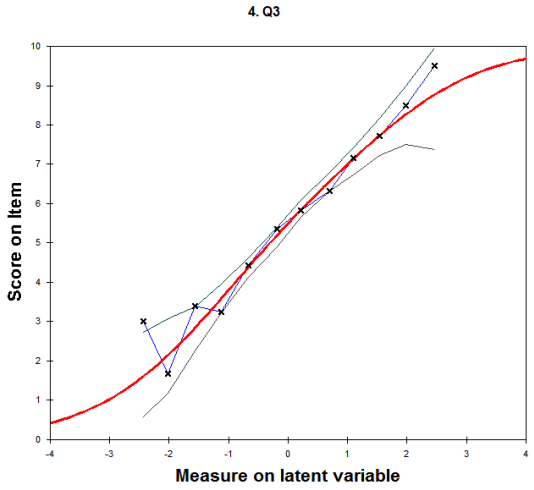


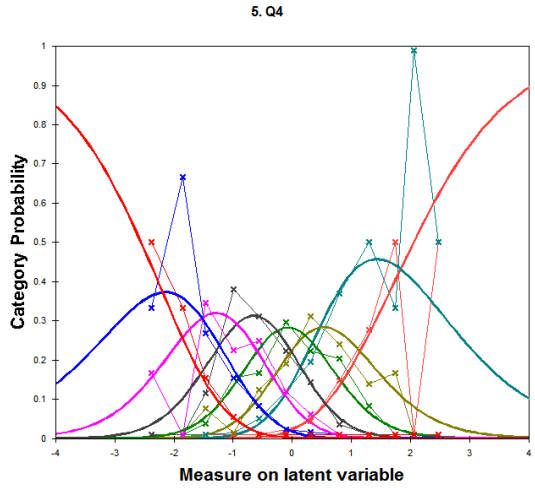


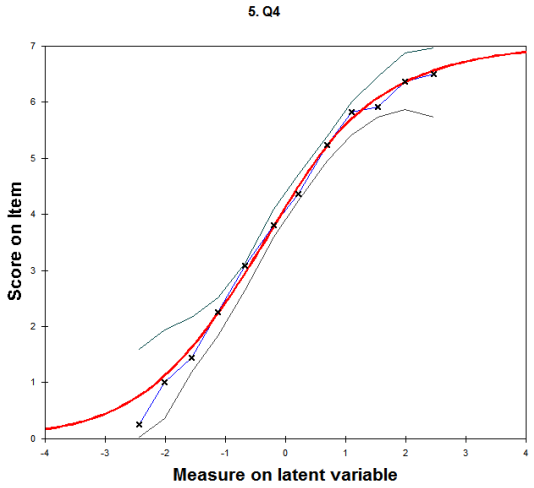


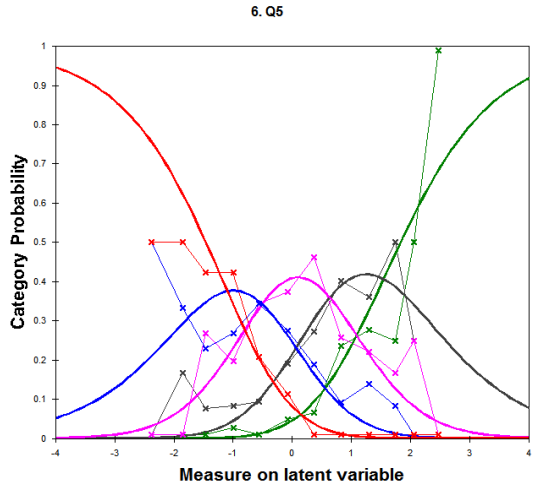


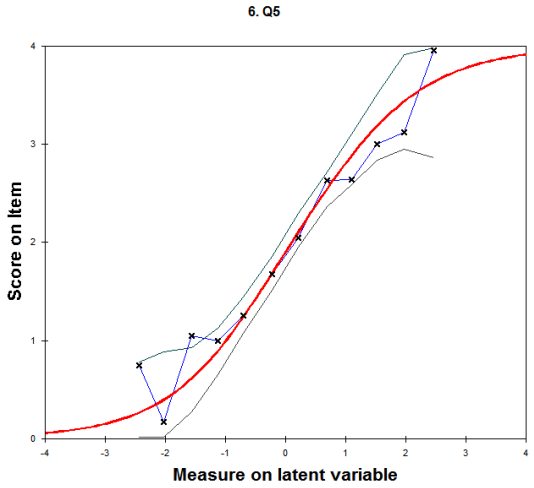


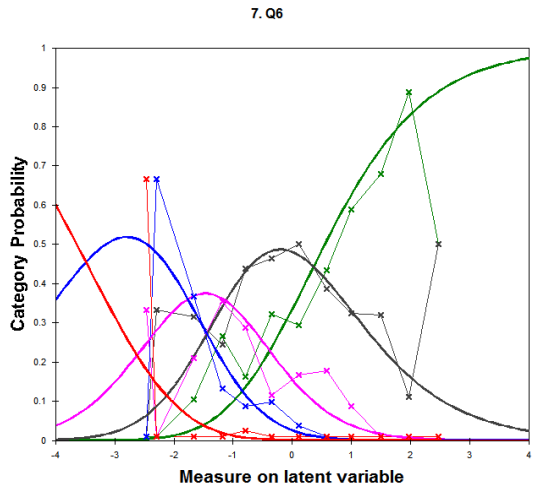


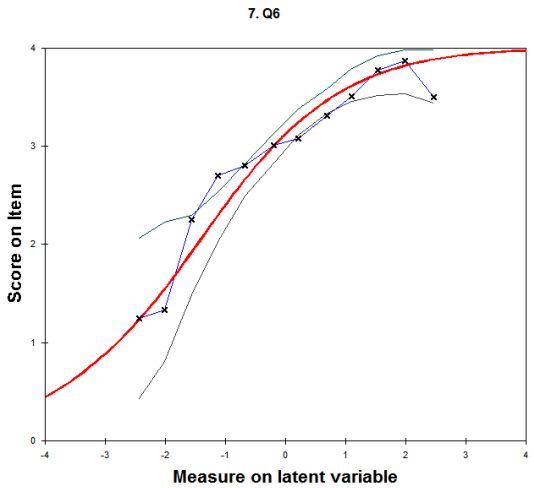

5. Q4
Category Probability
Measure on latent variable

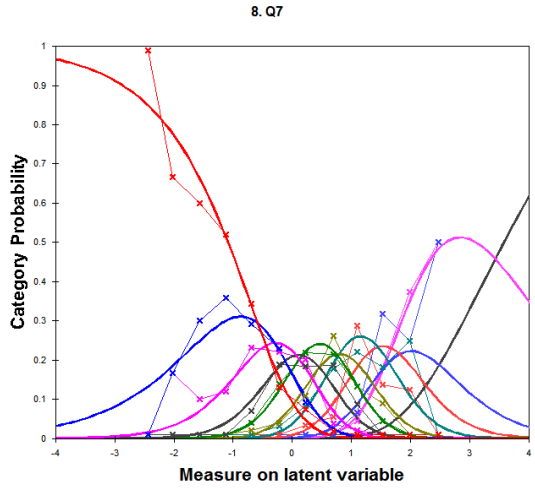

5. Q4
Score on Item
Measure on latent variable

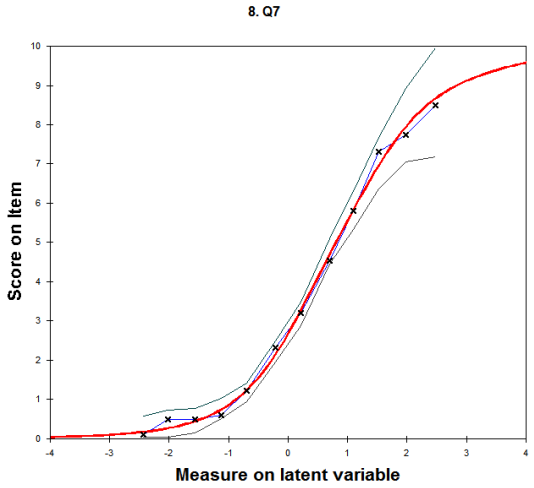

6. Q5
Category Probability
Measure on latent variable

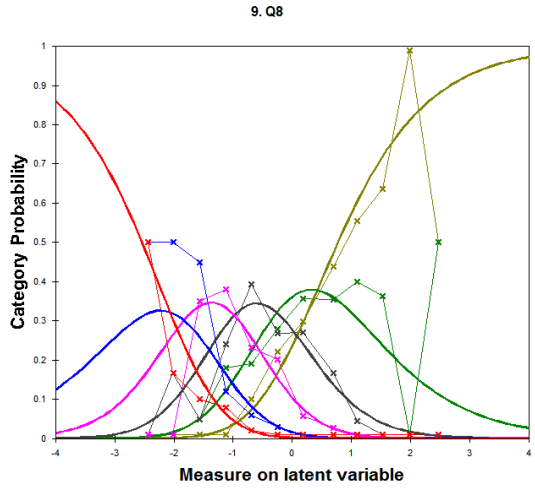

6. Q5
Score on Item
Measure on latent variable

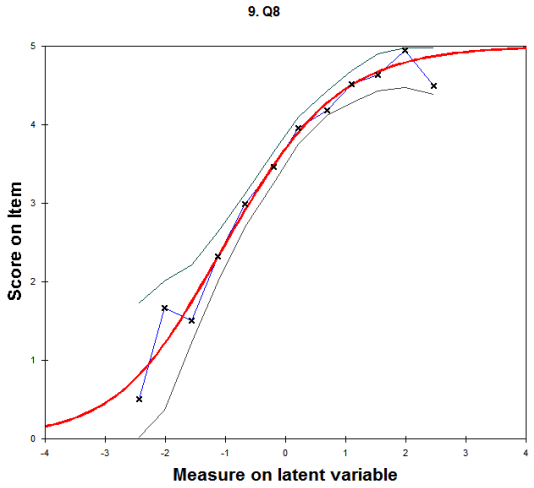

7. Q6
Category Probability
Measure on latent variable

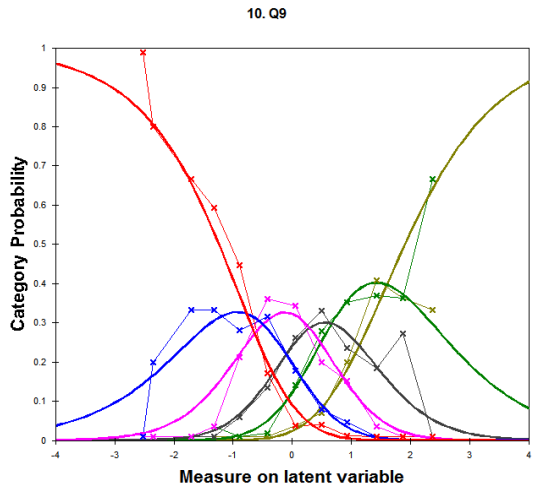

7. Q6
Score on Item
Measure on latent variable

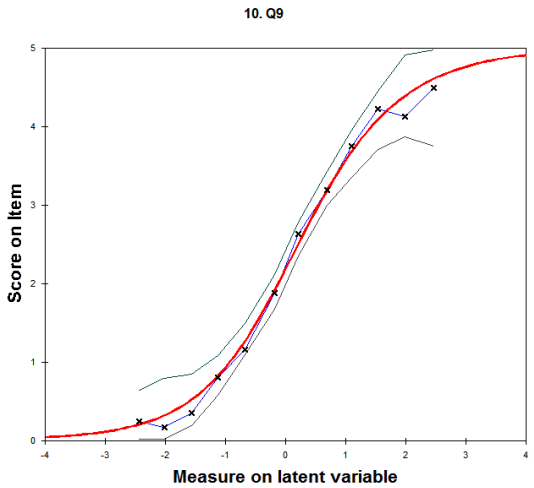

8. Q7
Category Probability
Measure on latent variable

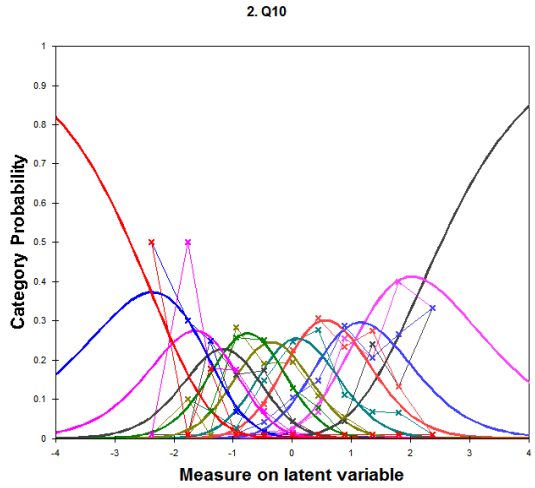

8. Q7
Score on Item
Measure on latent variable

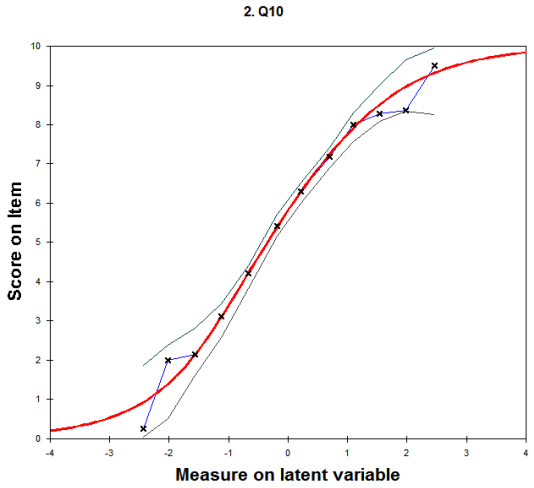

9. Q8
Category Probability
Measure on latent variable

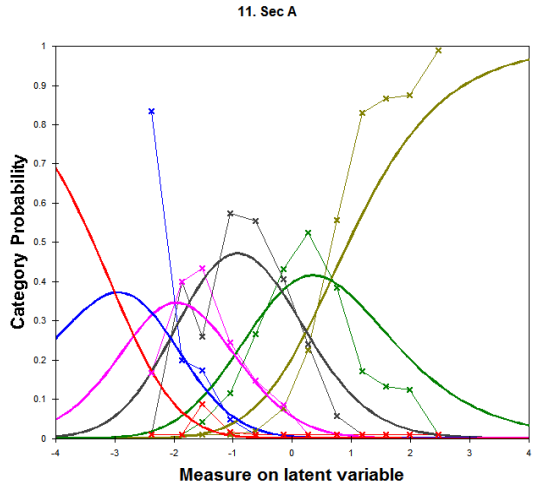

9. Q8
Score on Item
Measure on latent variable

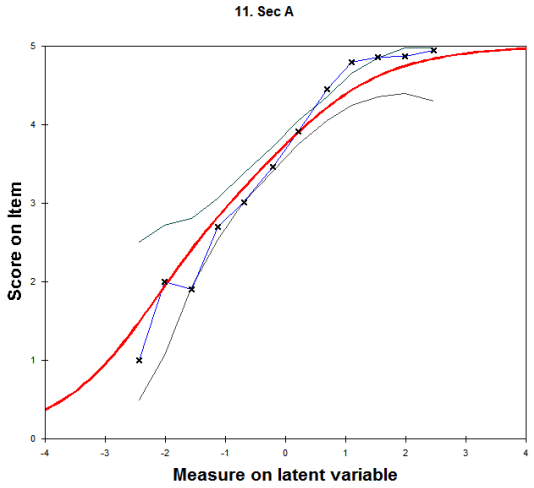

10. Q9
Category Probability
Measure on latent variable

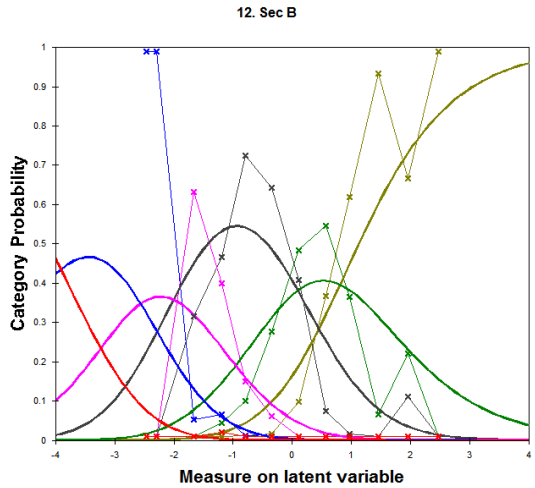

10. Q9
Score on Item
Measure on latent variable

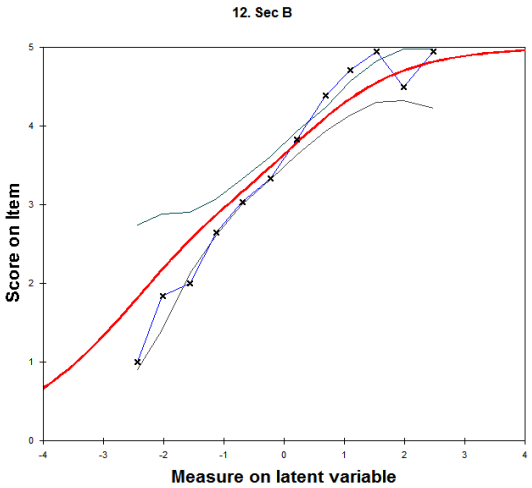

2. Q10
Category Probability
Measure on latent variable


2. Q10
Score on Item
Measure on latent variable


11. Sec A
Category Probability
Measure on latent variable

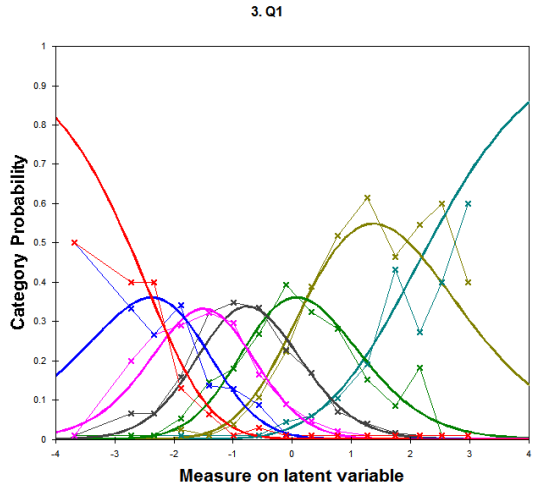

11. Sec A
Score on Item
Measure on latent variable

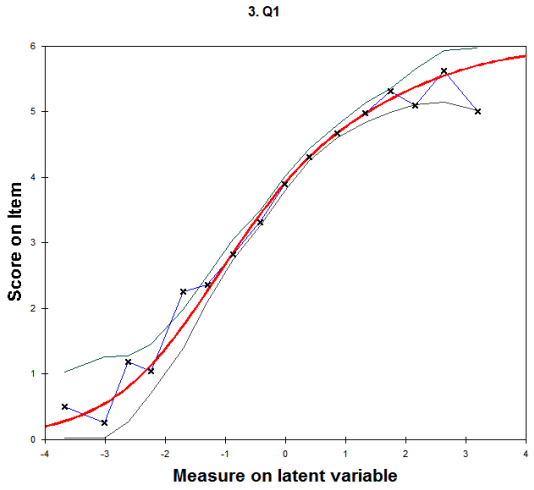

12. Sec B
Category Probability
Measure on latent variable

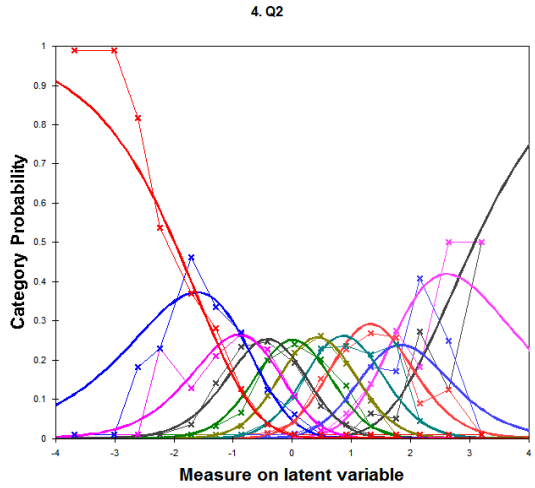

12. Sec B
Score on Item
Measure on latent variable

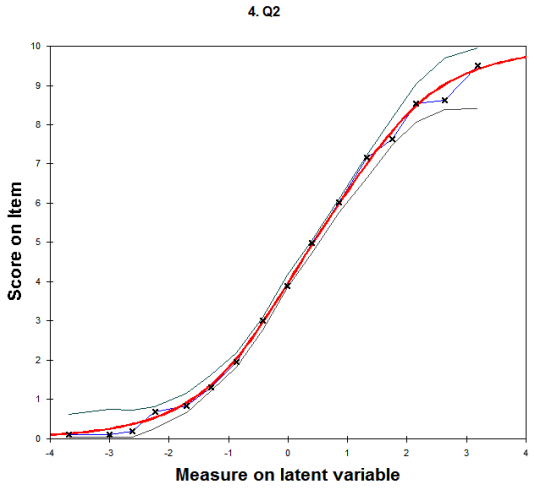
Appendix L

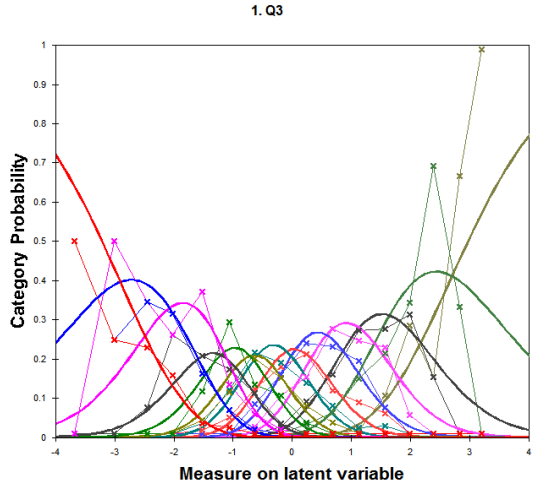
Item-level fit characteristics for WJEC's FN4 in June 2013

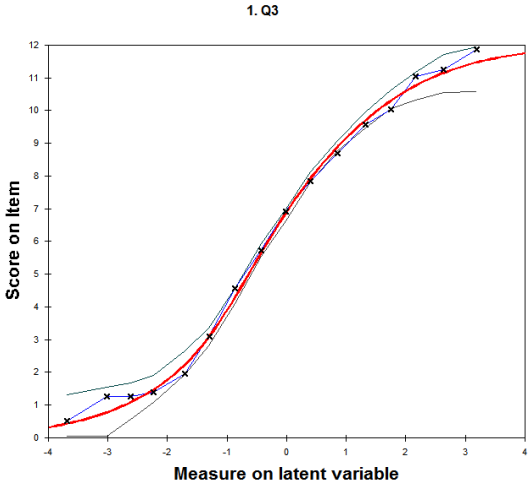


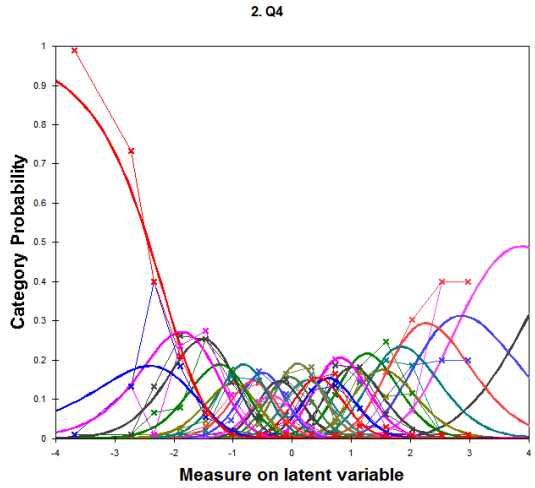


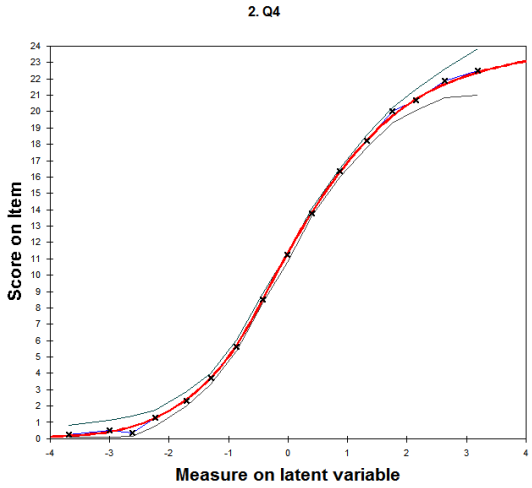






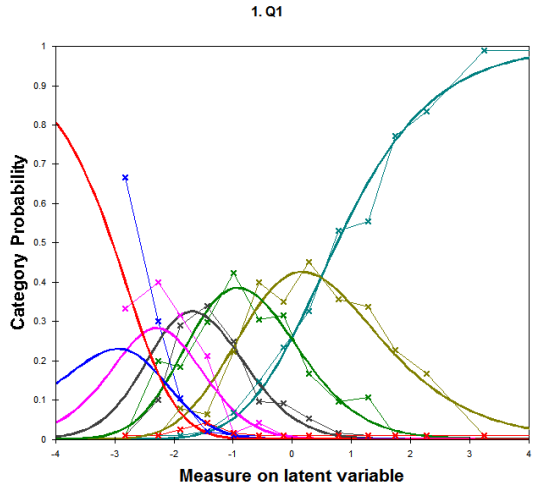


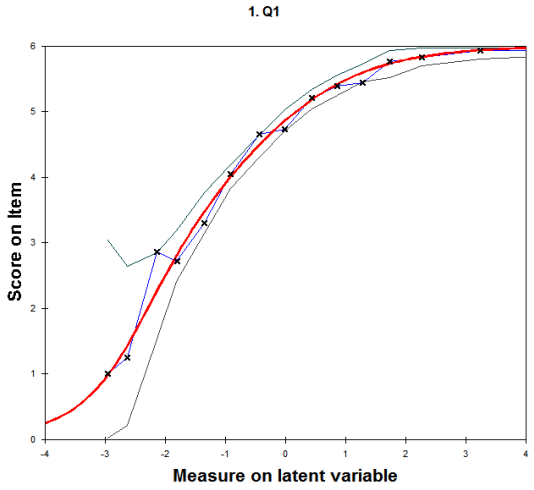

2. Q4
Category Probability
Measure on latent variable

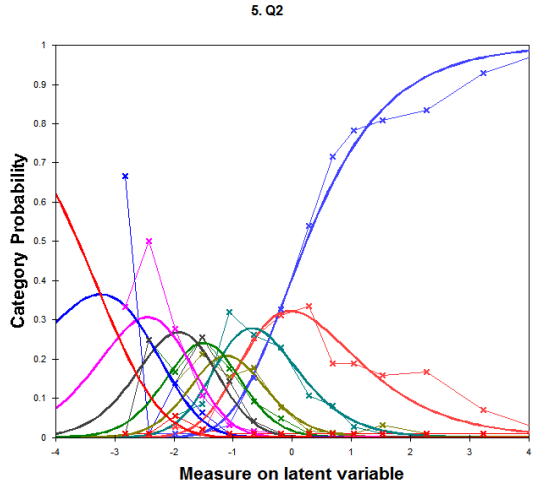

2. Q4
Score on Item
Measure on latent variable

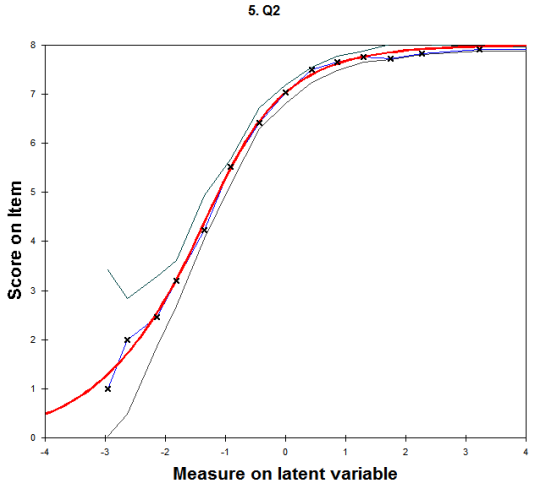
Appendix M

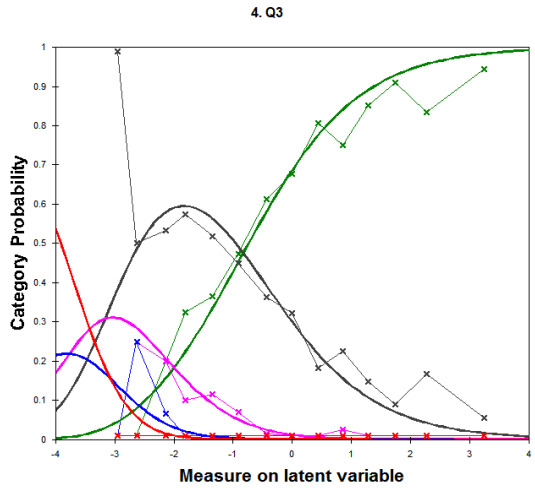
Item-level fit characteristics for WJEC's GN4 in June 2013

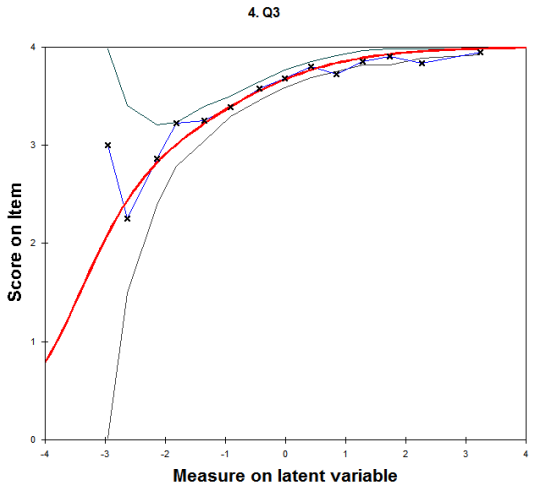


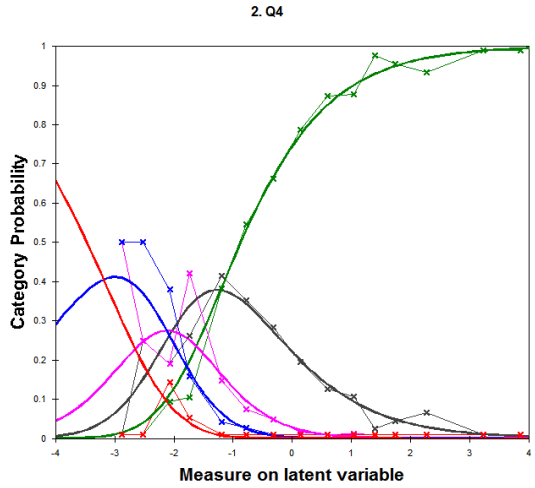


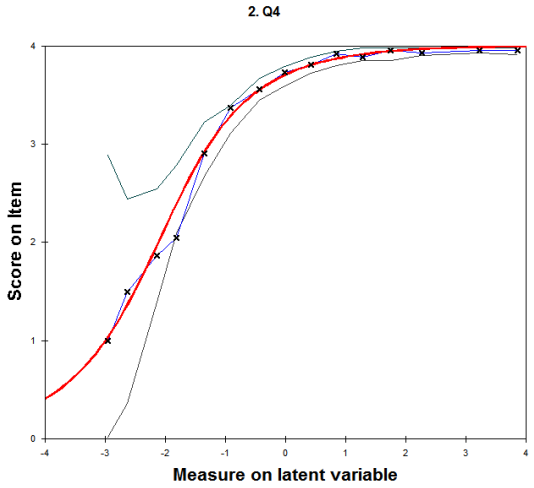


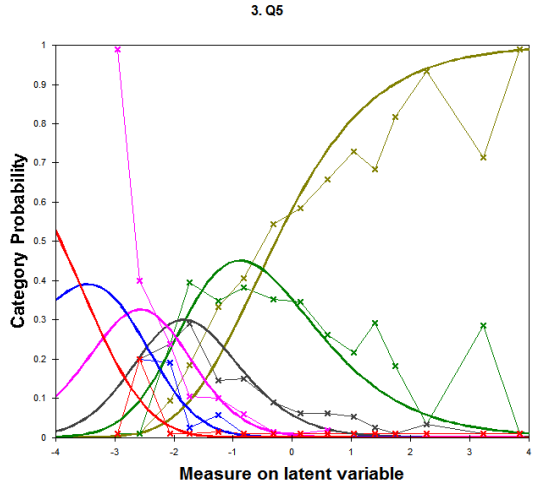


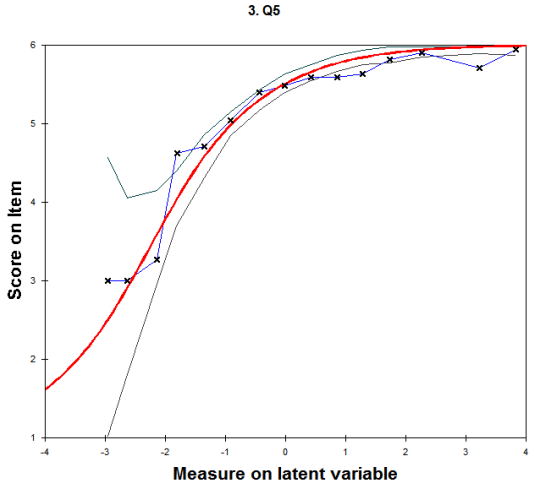


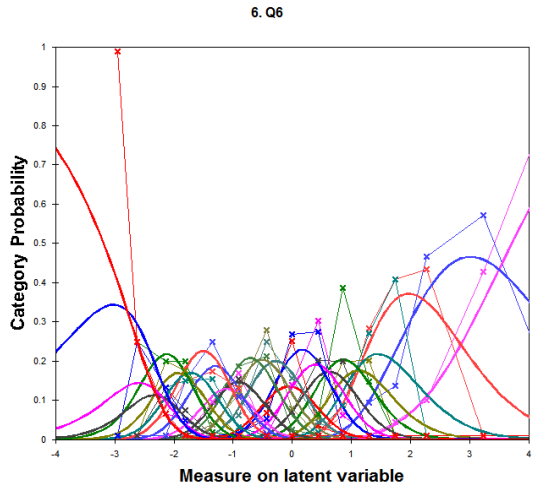


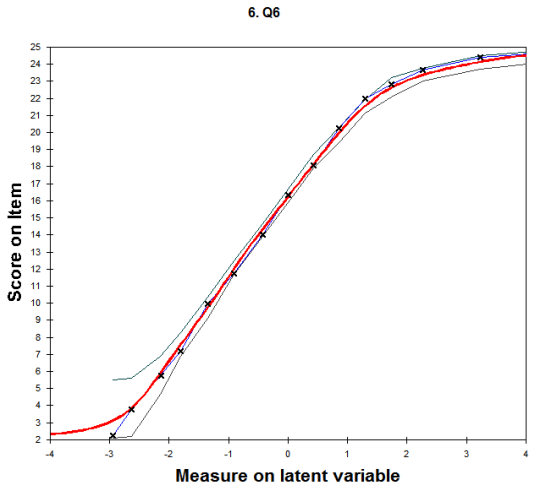

2. Q4
Category Probability
Measure on latent variable


2. Q4
Score on Item
Measure on latent variable


3. Q5
Category Probability
Measure on latent variable

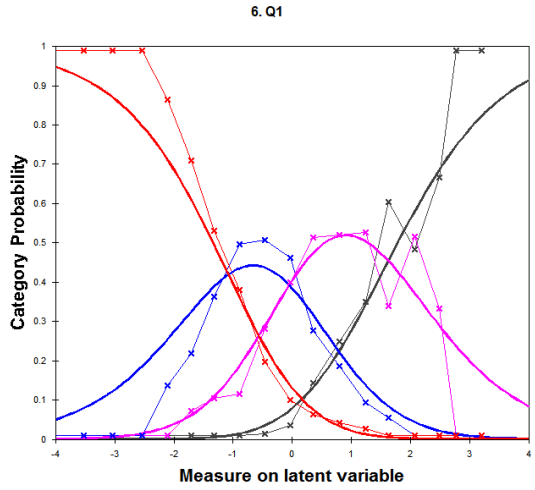

3. Q5
Score on Item
Measure on latent variable

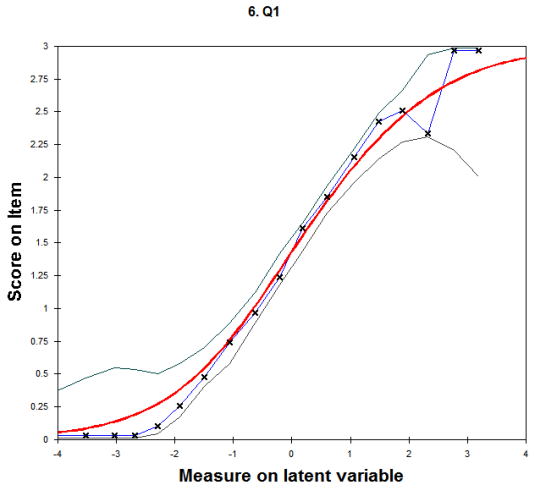

6. Q6
Category Probability
Measure on latent variable

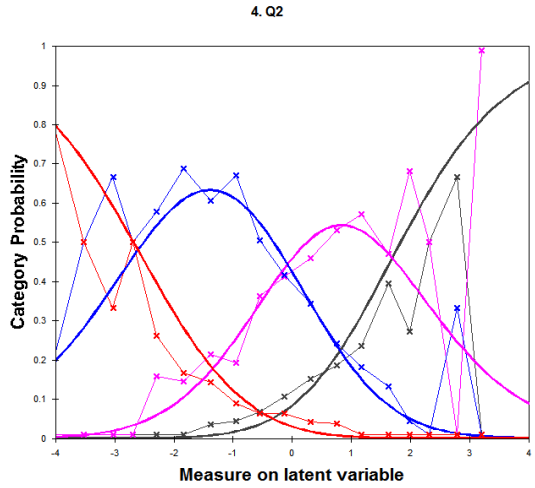

6. Q6
Score on Item
Measure on latent variable

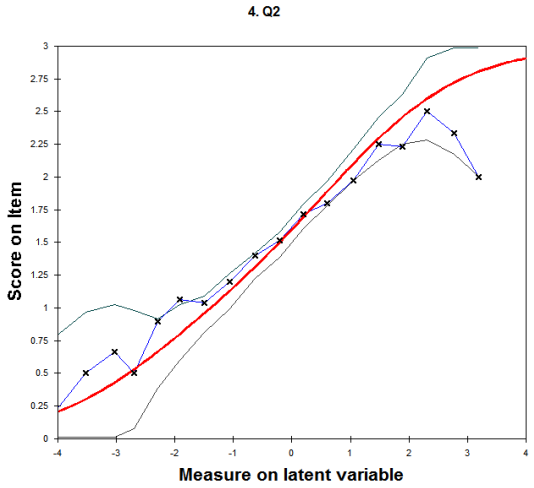
Appendix N

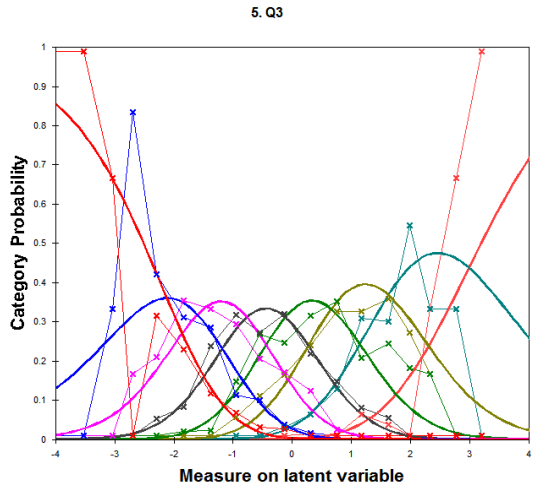
Item-level fit characteristics for WJEC's SN4 in June 2013

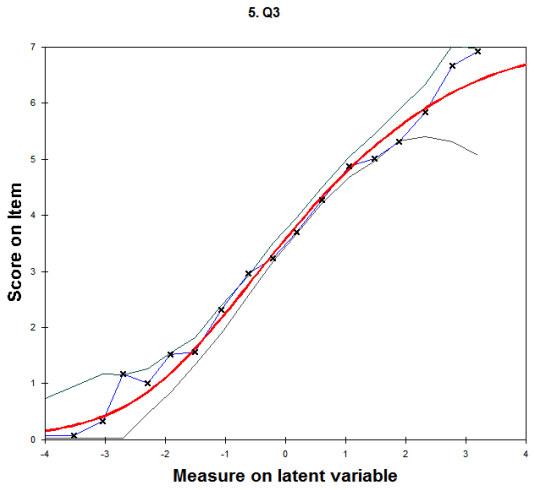


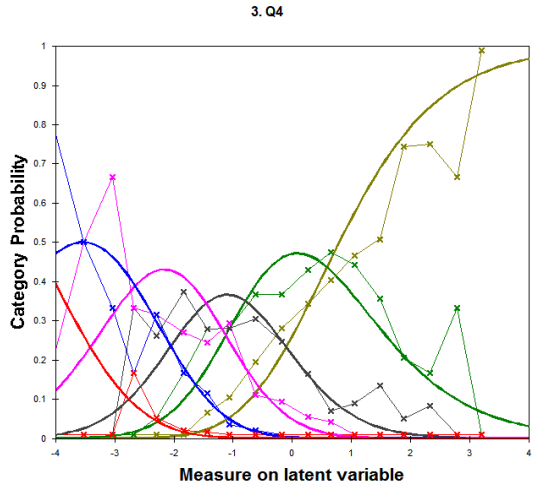


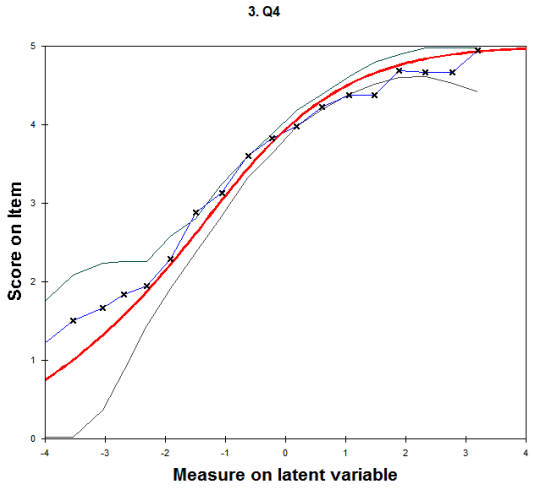


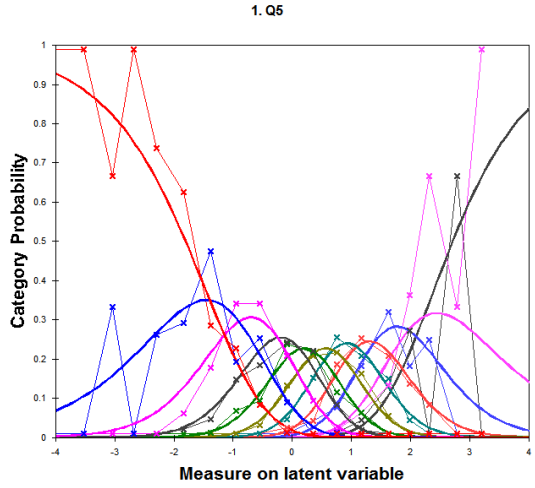


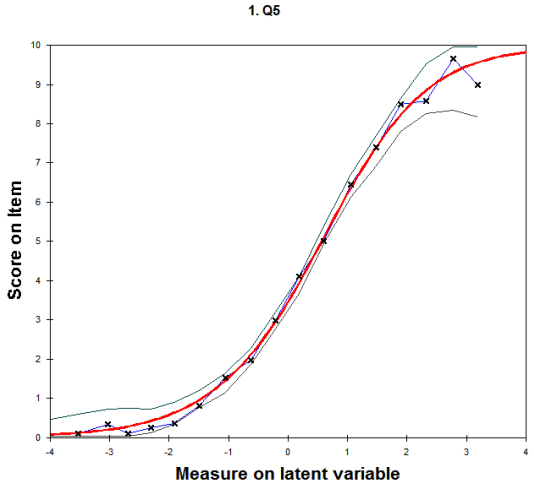


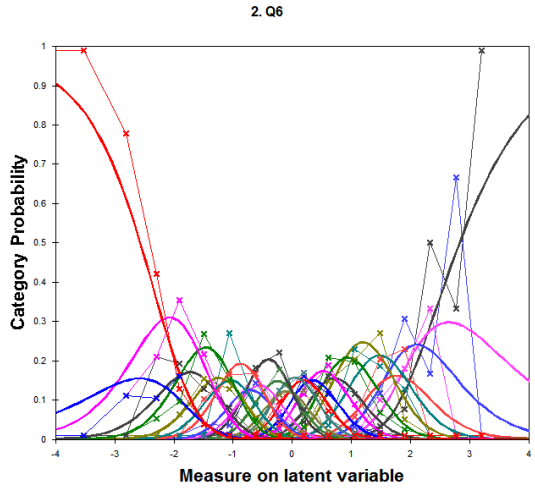


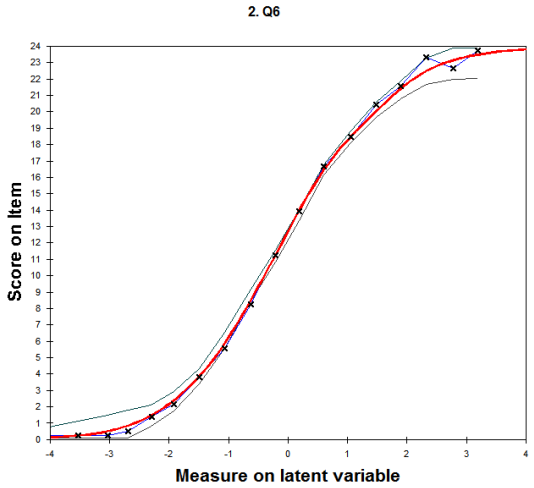

3. Q4
Category Probability
Measure on latent variable


3. Q4
Score on Item
Measure on latent variable


1. Q5
Category Probability
Measure on latent variable

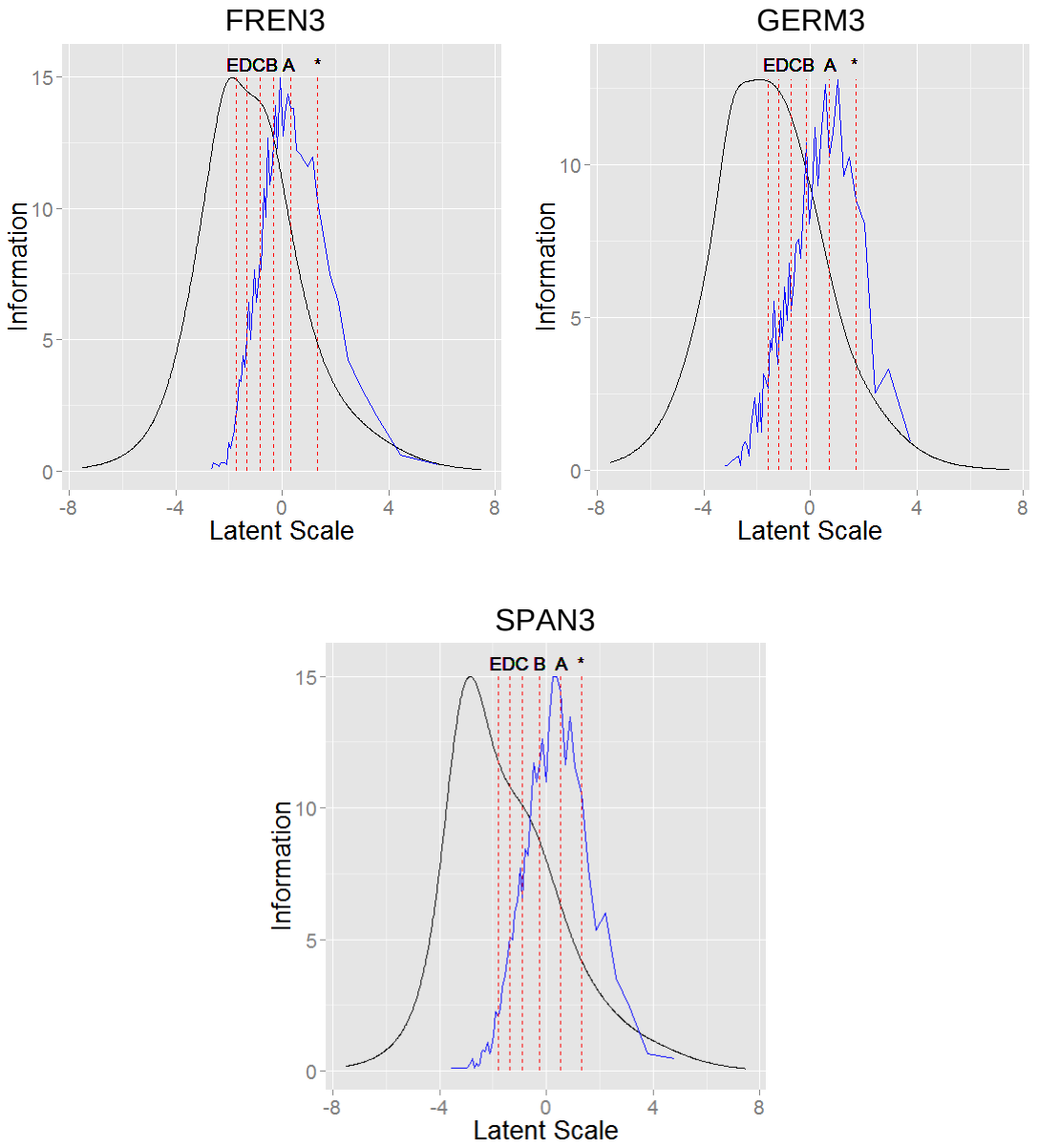

1. Q5
Score on Item
Measure on latent variable


2. Q6
Category Probability
Measure on latent variable


2. Q6
Score on Item
Measure on latent variable

Appendix O

Test information functions for section A of AQA's A2 written exams in summer 2013



We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

Published by the Office of Qualifications and Examinations Regulation in 2014

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346