

Psychometric and statistical analysis of the pilot delivery of English Level 1 externally-assessed achievement standards using digital medium

Dr Michael Johnston, Victoria University of Wellington

Eldon Paki, New Zealand Qualifications Authority

Background

This report presents statistical analyses comparing the psychometric properties of externally-assessed achievement standards for Level 1 English presented and completed in digital format with those of the same assessments presented and completed in paper format. The purpose of the analyses was to investigate the extent to which the two formats afforded candidates equivalent assessment opportunities. Such equivalence is important to establish before digital assessment is implemented on a large scale.

Executive summary

A number of analyses were undertaken to investigate the extent to which digital and paper formats for otherwise identical external assessments for Level 1 English were equivalent in respect of yielding results commensurate with the achievement levels of candidates.

In an initial analysis, the overall grade distributions from the digital-assessment format for each standard were compared with those from paper-based assessment format within the 45 participating schools. There were some statistically significant differences between the two sets of grade distributions. Specifically, for all of the Level 1 English standards in the digital pilot, percentages of *Not Achieved* results were significantly higher for paper-format assessments than for digital-format assessments, and percentages of *Merit* results and percentages of *Excellence* results were significantly and commensurately lower. In a complimentary analysis, the difficulty-parameter values from a Rasch analysis conducted on the digital-format data were compared with the difficulty parameters from a similar analysis conducted on the paper-format data. Again, some statistically significant differences between corresponding pairs of parameters from the digital and paper-based analysis were evident. The grade levels at which these differences occurred were generally consistent with the significant differences in the percentage-distributions of results.

In interpreting these differences it is important to consider that the groups of candidates undertaking assessments in each of the digital and paper formats were self-selecting and that there is therefore no basis to assume that the two groups were equal in ability, or that they ought to have attained the same distributions of results. To investigate the extent to which the differences between the digital and paper formats in results distributions and Rasch difficulty parameters were attributable to differences in the characteristics of the candidates undertaking the assessment in each format, rather than, or as well as, to characteristics of the assessment formats, a pair of linear regression analyses was conducted. These analyses modelled the relationship between internal assessment and external assessment in each of the digital and paper formats. The two regression models did not differ significantly in their parameter estimates, suggesting that the predictive relationship between internal and external assessment results was very similar for both the digital and paper formats. Thus, the differences in grade distributions and difficulty parameters are largely attributable to differences between the two groups of candidates; these analyses show no evidence that the format of the external assessment – digital or paper – affects the difficulty of the assessment.

In a second set of analyses an attempt was made to control for the probable difference in ability between the digital and paper-based groups evident in the first set of analyses by matching for performance on internally-assessed Level 1 English standards. These analyses were designed to determine whether there were any differences between the performance of digital-format candidates and paper-format candidates in externally-assessed achievement standards when differences in ability was controlled for. Any such residual differences could be attributed to characteristics of the two formats.

To attain the matched samples, a candidate who completed the external assessment in paper format was selected as a match for each candidate who completed the external assessment in digital format. To be considered a match, paper-format candidates had to have the same results in the same set of internally-assessed achievement standards as the digital-format candidate to whom he or she was to be matched. A paper-format candidate was randomly selected from the set of all candidates who were matches for each digital-format candidate.

Although the distributions of results for the digital- and paper-format candidates were much more closely matched than they were in the first set of analyses, some statistically significant differences remained. Specifically, for all of the Level 1 English standards in the digital pilot, percentages of *Not Achieved* results were significantly higher for paper-format assessments than for the digital-format assessments. Rasch analyses also showed significant differences between the difficulty parameters estimated for each format, consistent with the differences in the percentage-distributions of results. This finding demonstrates small differences in favour of the digital assessment format at the level of gaining credit, but not for higher grades.

Taken together, the regression and matched-sample results suggest that both the characteristics of the sets of the candidates undertaking external assessment in each format and the characteristics of the formats themselves, are influential on the observed differences in the overall results distributions. Importantly, the latter seem to be in favour of the digital format, suggesting that, if digital assessment was to be implemented on a large scale, some rise in the proportions of students gaining credits in external assessment, at least in assessments with similar characteristics to the Level 1 English assessments investigated here, might be expected.

Sample characteristics

A total of 45 schools took part in the pilot of digital examinations for the three externally-assessed achievement standards for Level 1 English in 2017. The data supporting these analyses were the externally-assessed Level 1 English results for all candidates from these schools.

For standard 90849, 43 schools contributed results from the digital format and 40 from the paper-based format. For standard 90850, 42 and 41 schools contributed results from the digital and paper-based formats respectively. For standard 90851, 33 schools contributed digital-format results and 34 contributed paper-format results. A grand total of 14,147 results, 43.2 percent of which were from the digital examination format, were collected for the pilot from 2,759 candidates.

Table 1 shows the number and name of each externally-assessed achievement standard for Level 1 English, as well as the total numbers of results for each of the paper and digital formats at the participating schools, and the proportions of all results that were for candidates completing each standard in digital format. Similar percentages (between 41% and 46%) of candidates for each standard used digital format. The standard with the fewest results overall – *Show understanding of significant aspects of unfamiliar written text(s) through close reading, using supporting evidence* (90851) – had the greatest proportion of candidates undertaking it in digital format.

Table 1.

Total numbers of results for Level 1 externally-assessed achievement standards in English at the 45 schools participating in the digital assessment pilot.

Standard Number	Standard Title	Total results: Digital format	Total results: Paper format	Digital format results as percentages of all results (%)
90849	Show understanding of specified aspect(s) of studied written text(s), using supporting evidence	2,113	2,983	41.5
90850	Show understanding of specified aspect(s) of studied visual or oral text(s), using supporting evidence	2,239	2,956	43.1
90851	Show understanding of significant aspects of unfamiliar written text(s) through close reading, using supporting evidence	1,759	2,097	45.6
Total		6,111	8,036	43.2

Comparison of overall grade distributions for digital and paper formats

Figure 1 compares the distributions of grades for digital and paper formats for each of the three standards included in the pilot. Individual results distributions for each standard at each participating school, from each of the digital and paper formats, are shown in the appendix.

For all three standards, the percentages of both *Not Achieved* and *Achieved* grades were higher for paper format than for digital format, and the percentages of *Merit* and *Excellence* grades were commensurately lower. *Z* tests showed that the differences between the two formats in the percentages of *Not Achieved*, *Achieved*, *Merit* and *Excellence* grades were statistically significant ($p < .05$) for

standards 90849 and 90850. For standard 90851, the percentages of *Not Achieved*, *Merit* and *Excellence* grades differed significantly between the two formats, but not the percentages of *Achieved* grades. The magnitudes of the differences between the formats in the percentages of candidates gaining credit ranged from 2.4 percentage points in 90849 to 5.5 percentage points in 90851.

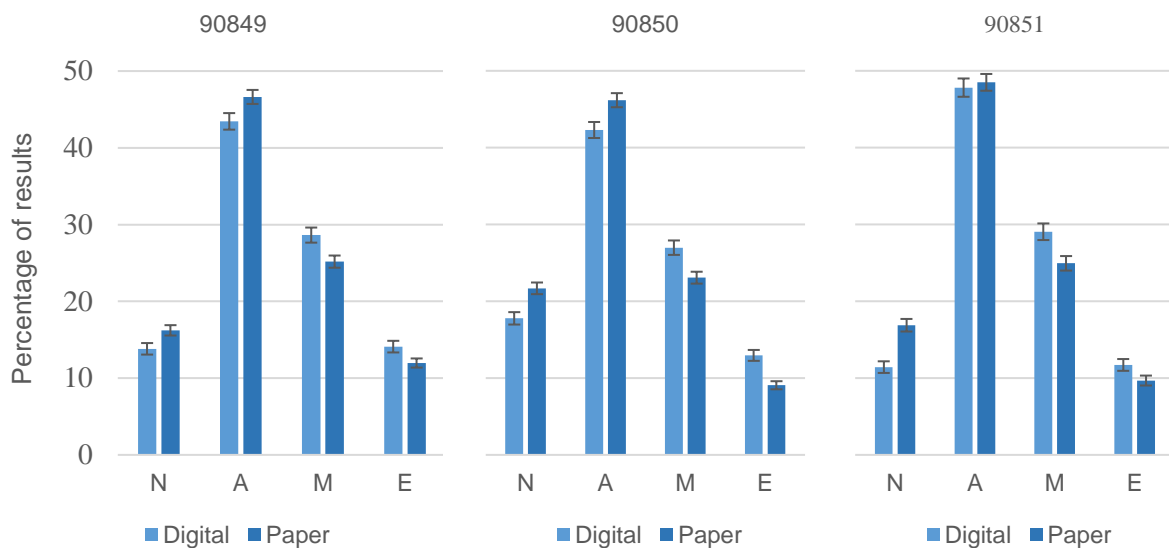


Figure 1. Comparisons of digital and paper results distributions for Level 1 externally-assessed achievement standards in English at schools participating in the digital pilots. Vertical bars denote standard errors.

Candidates were not randomly allocated to digital- or paper-format assessments, but selected which format to undertake themselves. Therefore the observed differences in the results distributions for the two formats might be attributable to differences in the capabilities of candidates completing assessments in each of the two formats rather than, or as well as, to differences in the characteristics of the two assessment formats. In particular, more capable candidates might, on average, have felt more confident to use the digital format than less capable candidates, which would at least partly explain the difference in the results distributions in favour of the digital format. Thus, the difference in the distributions for the two formats does not, on its own, provide compelling evidence that the two formats differ in respect of their accessibility to candidates.

Comparison of Rasch difficulty parameters for digital and paper formats

Subsets of candidates who completed assessments *only* in digital format and *only* in paper format were identified. Three Rasch analyses were conducted for each of these subsets to estimate difficulty parameters associated with grades of *Achieved* or better, grades of *Merit* or *Excellence* and grades of *Excellence*. In each of these analyses, the individual standards were treated as items, so the measurement scale on which the difficulty parameters were estimated reflected aggregated performance across the three standards, allowing for a quantitative (interval scale) comparison of difficulty. Figure 2 shows comparisons of these parameters for each of the three standards.

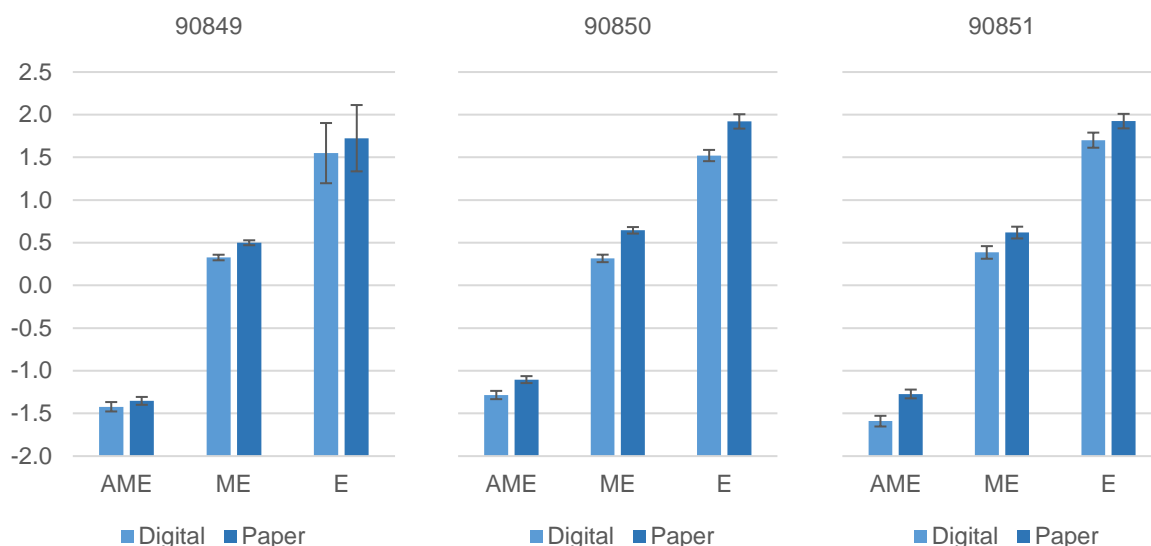


Figure 2. Comparisons of Rasch difficulty parameters for attaining grades of *Achieved or better* (AME), *Merit or better* (ME) and *Excellence* (E), for digital and paper formats for each externally-assessed Level 1 achievement standard in English. Vertical bars denote standard errors.

For all three standards the difficulty parameters associated with attaining grades of *Achieved or better*, with attaining grades of *Merit or better* and with grades of *Excellence* were higher for the paper format than for the digital format, meaning that candidates found the paper format more difficult than the digital format. The differences between the pairs of parameter values for the two formats exceeded the 95% confidence interval for the difference in the case of ME for standard 90849, for AME, ME and E for standard 90850 and for AME for standard 90851. These differences may therefore be treated as being statistically significant with $p < .05$. No other differences between pairs of parameters exceeded the 95% confidence interval for the difference.

These analyses largely corroborate the analyses of differences in grade distributions depicted in Figure 1, and are subject to the same caveat that the differences may be attributable to characteristics of the groups of candidates undertaking the assessment in each format rather than – or in addition to – characteristics of the formats themselves. In other words, the candidates may have found the paper format more difficult because they were, on average, less able than the candidates completing the assessment in digital format, because the paper format has characteristics that made achievement less accessible, or for some combination of these reasons.

Analysis of the predictive relationship between internal assessment and external assessment in each of the digital and paper formats

To investigate the source of the differences in grade distributions (Figure 1) and Rasch difficulty parameters (Figure 2) – that is, whether they are attributable to characteristics of candidates or characteristics of the formats – the equivalence of the digital and paper mediums in terms of the extent to the level of performance in each format predicted by a given level of performance in the *internally-assessed* achievement standards for Level 1 English was analysed.

The performance variables for the two external assessment formats comprised the ability estimates for each candidate resulting from the Rasch analysis described above. A third Rasch analysis was carried out on all internally-assessed results of participating candidates – both those who completed all of the external assessments digitally and those who completed all of them on paper. Like the analyses of the two external assessment formats, this analysis treated individual internally-assessed standards as items, yielding an interval-scale measurement variable as an aggregate measure of performance across internally-assessed standards.

A least-squares linear regression analysis was used to model the relationship between digital-format external assessment and internal assessment and another to model the relationship between paper-format external assessment and internal assessment. Figure 3 depicts two scatterplots, with regression lines, showing the relationship between digital-format external assessment and internal assessment (upper panel) and paper-format external assessment and internal assessment (lower panel).

A comparison of the constant and slope parameters estimated under regression models for each external assessment format allows for a statistical comparison of the equivalence of the predictive relationships. These comparisons showed that the differences between both the constants and the slope parameters were well inside the 95% confidence intervals for their respective differences. Thus there is no evidence arising from this analysis that the two formats differed in respect of the predictive relationship between internal and external assessment. This finding suggests that it is characteristics of the groups of candidates, rather than characteristics of the digital or paper assessment formats, that explains the differences in grade distributions and Rasch difficulty parameters depicted in Figures 1 and 2 respectively.

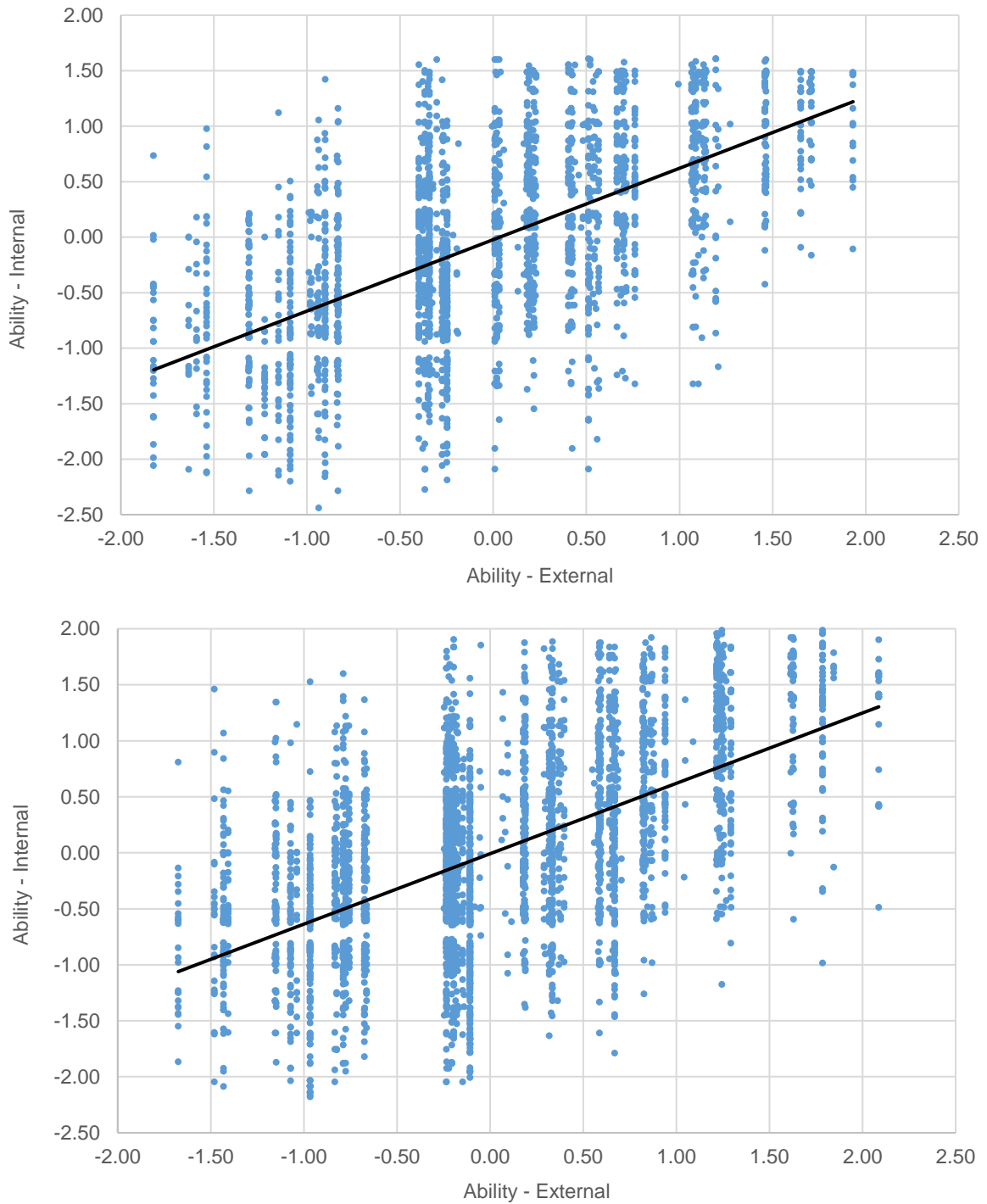


Figure 3. Scatterplots with regression lines showing the relationships between Rasch ability parameters estimated for external assessment and internal assessment. The upper panel shows the relationship for digital-format external assessment and the lower panel the relationship for paper-format external assessment.

Comparison of grade distributions for digital and paper formats on matched samples

To further investigate the source of the discrepancies between the digital and paper-format results shown in Figures 1 and 2, a sample of candidates was selected from those completing external assessments in the paper-based format, matching each digital-format candidate for their results in internal assessment for Level 1 English. For each digital-format candidate, a matching paper-format candidate was randomly selected from the set of all paper-format candidates with the same profile of internal assessment results – that is, from those paper-format candidates who undertook the same set of internally-assessed standards and attained the same result for each – as the target digital-format candidate. This approach, to some extent at least, controls for differences in ability between the sets of candidates completing the external assessment in each format.

Figure 4 shows the grade distribution for the digital-format candidates and for the matched sample paper-format sample, for each of the three externally-assessed standards. For all three, the percentage of *Not Achieved* grades was higher for paper format than for digital format, and the percentages of *Merit* and *Excellence* grades were lower. Z-tests showed that the percentages of *Not Achieved* grades for paper format were significantly higher ($p < .05$) than for digital format for all three standards and that the percentages of *Merit* grades for paper format for standard 90851 was significantly lower ($p < .05$) than for the digital format. No other comparisons of the same grades for the same standards differed significantly between the two samples. The magnitudes of the differences in percentages of candidates gaining credit ranged were 2.4 percentage points in 90849 and 90850 and 5.3 percentage points in 90851. These magnitudes were very comparable with those found in the comparisons of overall grade distributions (Figure 1), although all but one of the differences in percentages of candidates attaining higher grades evident in the overall comparison were eliminated in the match-sample comparison.

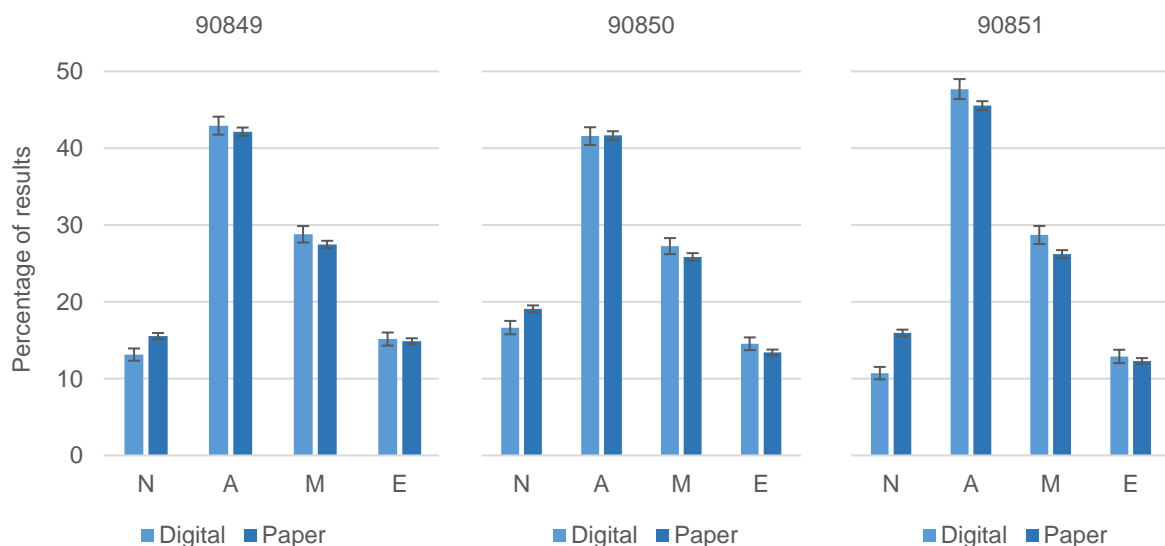


Figure 4. Comparisons of digital and paper results distributions of the matched samples for Level 1 externally-assessed achievement standards in English. Vertical bars denote standard errors.

Comparison of Rasch difficulty parameters for digital and paper formats on matched samples

In a final analysis, difficulty parameters associated with grades of *Achieved* or better, grades of *Merit* or *Excellence* and grades of *Excellence* for each externally-assessed standard were estimated using

separate Rasch analyses of the data from the digital-format candidates and the matched paper-format candidates. As was the case for the analysis depicted in Figure 2, in each of these analyses, the individual standards were treated as items, so that the measurement scale on which the difficulty parameters were estimated reflected aggregated performance across the three standards, allowing for a quantitative (interval scale) comparison of difficulty. Figure 5 shows comparisons of these parameters for each of the three standards.

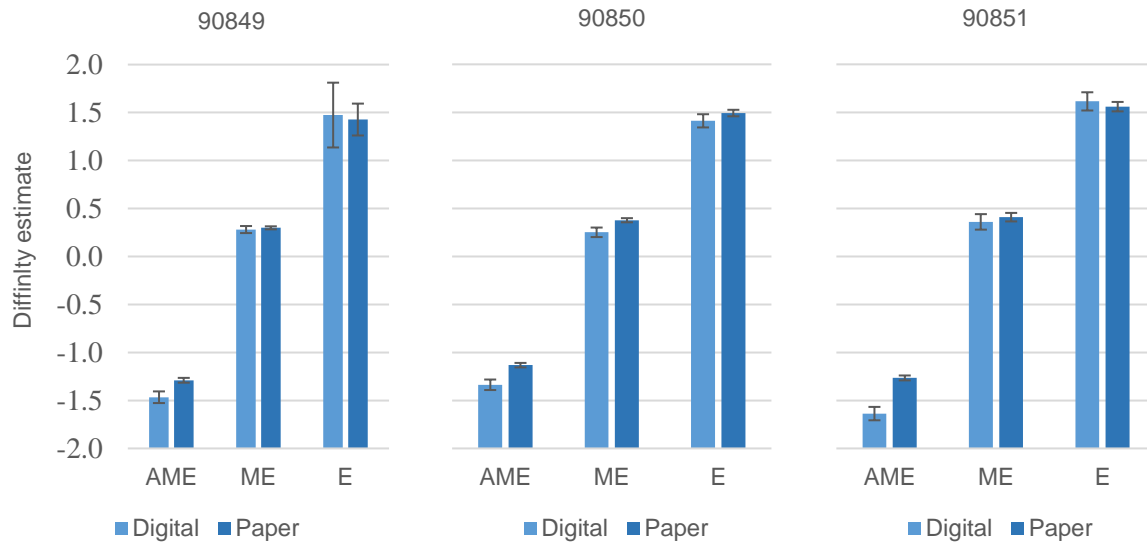


Figure 5. Comparisons of Rasch difficulty parameters for attaining grades of *Achieved* or better (AME), *Merit* or better (ME) and *Excellence* (E), of the matched samples for each externally-assessed Level 1 achievement standard in English. Vertical bars denote standard errors.

For all three standards the difficulty parameters associated with attaining grades of *Achieved* or better and with attaining grades of *Merit* or better were higher for the paper format than for the digital format. The differences between the pairs of parameter values for the two formats exceeded the 95% confidence interval for the difference in the case of AME for all three standards and for ME for standard 90850. These differences may therefore be treated as being statistically significant. No other differences between corresponding pairs of parameters exceeded the 95% confidence interval for the difference.

Discussion

The analyses of the overall results distributions show some differences between those from the digital format and those from the paper-based format. In particular, performance in all three standards included in the pilot was somewhat better in the digital format than in the paper format. These differences were largely corroborated by corresponding differences in the difficulty-parameter estimates from the Rasch analysis of the overall data.

Least-squares regression analyses were used to investigate the predictive relationship between internal assessment and external assessment, with separate analyses conducted for each of the digital and

paper-based formats for the external assessment. These analyses showed no significant difference between these predictive relationships. In other words, according to these regression models, a candidate with a given level of attainment in internal assessment is predicted to achieve the same level attainment in external assessment, irrespective of whether that external assessment is conducted in a digital or paper-based format. Thus, the regression analysis suggests that the characteristics of the assessment format do not influence achievement and that, therefore, the group of candidates who elected to undertake external assessment in digital format were, on average, more able than those who elected to undertake it in paper format. A possible reason for this is that more able students are often also more confident, and that being a new approach to sitting examinations, more confident students were more likely to opt for the digital format than less confident students.

The analyses of the matched-sample data did not, however, entirely corroborate the regression analyses. While most of the differences between assessments formats in the grade distributions evident in the full data set were not present in the matched-sample data, differences at the level of gaining credit remained. These differences were again corroborated by comparisons of difficulty parameters in a corresponding Rasch analysis of the matched-sample data.

Taken together, the regression and matched sample analyses suggest that the characteristics of the candidates understanding assessment in the two formats and the characteristics of the format themselves are both associated with the higher performance observed in the digital format. As noted above, it is likely that more confident candidates have opted for the digital format, on the basis that less confident candidates might be more likely to want to use a testing format that is familiar to them. Under the assumption that confidence is correlated with ability, this would explain the influence of candidate characteristics on the observed difference. On the other hand, the matched-sample comparisons suggest that students, particularly those performing near the boundary of *Not Achieved* and *Achieved*, actually found the digital format more accessible.

Speculatively, a reason for the apparently-greater accessibility of credits in the digital format might be related to the fact that the assessments for Level 1 English are relatively writing intensive. One possibility in this regard is that students are, on average, more fluent with typing than with handwriting, especially under time pressure. Another is that, when marking scripts completed in digital format, markers do not have to contend with handwriting legibility issues.

In any event, the present analyses suggest that, at least in assessments requiring similar skills to Level 1 English, if digital assessment was to be adopted on a large scale, some increase in the proportions of students gaining credit might be expected.

APPENDIX

Numbers of results in each grade category for each participating school disaggregated by assessment format

School	Standard 90849									
	Digital					Paper				
	%N	%A	%M	%E	<i>n</i>	%N	%A	%M	%E	<i>n</i>
1	100	0	0	0	1					
2	36	55	9	0	11	50	36	14	0	22
3	12	60	20	7	108	6	63	26	5	140
4					0	16	44	23	17	100
5	13	29	38	21	24	19	44	27	10	150
6	49	31	4	16	55	57	14	0	29	7
7	23	55	18	5	22	14	42	31	12	182
8	4	41	40	15	196	7	34	33	26	203
9	9	22	34	34	58	2	28	43	27	89
10	14	43	24	18	283	24	44	20	11	45
11	8	47	31	14	290	24	38	13	24	45
12	38	52	5	5	21	24	60	10	6	50
13	20	56	20	4	55	31	60	10	0	42
14	67	33	0	0	3					0
15	0	25	25	50	4	16	41	29	14	143
16	23	64	14	0	22	30	49	18	2	89
17	5	68	27	0	22	21	45	25	9	104
18	3	45	33	19	58	14	58	22	6	252
19	20	60	0	20	5	60	40	0	0	5
20	75	25	0	0	8	100	0	0	0	2
21	13	50	25	13	8	8	63	21	8	52
22	17	50	33	0	18	100	0	0	0	2
23	100	0	0	0	3	18	58	17	7	71
24	5	51	24	19	37	9	50	28	14	58
25	10	28	45	18	40	13	35	38	14	71
26	30	55	15	0	20					0
27	11	48	33	7	27	18	52	24	6	116
28	17	55	23	5	86	50	50	0	0	2
29	27	53	20	0	15	24	44	21	11	211
30	100	0	0	0	8					0
31	100	0	0	0	1	17	57	12	14	127
32					0	15	77	8	0	13
33	34	49	15	2	47	13	47	34	6	123
34	1	30	42	27	168	0	24	31	45	29
35	18	55	27	0	11	26	52	19	3	69
36	14	37	33	16	116	21	42	32	5	19
37	8	33	43	16	51	12	38	35	15	106
38	4	52	32	12	25	6	49	31	14	51
39	19	58	18	4	67	21	41	23	15	92
40	13	39	37	11	70	67	33	0	0	3
41	27	36	36	0	11	14	66	21	0	29
42	0	17	17	67	6	8	15	23	54	13
43	6	38	44	13	16	9	34	34	23	47
44	29	43	29	0	7	11	89	0	0	9
45	89	11	0	0	9					0
Total	14	43	29	14	2,113	16	47	25	12	2,983

School	Standard 90850									
	Digital					Paper				
	%N	%A	%M	%E	<i>n</i>	%N	%A	%M	%E	<i>n</i>
1	100	0	0	0	1					0
2	70	30	0	0	10	59	23	18	0	22
3	7	47	37	10	90	11	57	24	7	87
4	41	41	18	0	17	17	43	35	5	162
5	22	15	22	41	27	22	39	29	10	157
6	74	19	5	2	42	67	33	0	0	6
7	5	77	9	9	22	21	46	26	7	173
8	5	36	42	18	192	10	35	36	19	155
9	3	16	53	28	58	2	31	46	20	89
10	20	40	25	15	302	34	39	15	11	61
11	12	38	30	20	288	25	41	23	11	44
12	50	45	5	0	22	26	60	14	0	50
13	30	42	21	8	77	39	48	8	5	62
14	18	69	10	3	39	0	100	0	0	1
15	0	25	25	50	4	9	44	32	15	154
16	27	45	27	0	11	51	37	10	2	94
17					0	43	46	11	0	35
18	12	29	47	12	58	25	59	13	3	304
19	50	50	0	0	6	40	50	10	0	10
20	50	50	0	0	8					0
21	0	56	33	11	9	21	52	21	6	52
22	29	43	24	5	21	62	38	0	0	13
23	55	45	0	0	31	24	56	17	3	117
24	13	38	28	21	47	30	35	25	10	40
25	10	36	26	29	42	18	49	17	16	125
26	43	52	5	0	21					0
27	4	44	41	11	27	15	46	27	13	128
28	22	58	18	1	98	33	67	0	0	3
29					0	6	42	31	22	36
30	23	54	15	8	39	0	50	0	50	2
31	15	65	17	2	52	22	47	22	9	143
32	67	33	0	0	3	56	44	0	0	16
33	27	58	13	2	48	16	59	22	3	79
34	13	43	35	9	131	15	61	20	5	41
35					0					0
36	18	50	21	12	121	52	24	14	10	21
37	12	35	27	27	52	14	46	22	18	103
38	5	55	32	9	22	14	43	24	18	49
39	11	45	33	11	66	41	32	23	3	115
40	7	35	41	17	71	25	50	25	0	4
41	0	100	0	0	18	23	61	13	3	61
42	13	44	31	13	16	0	17	58	25	12
43	25	25	25	25	4	12	46	27	15	119
44	43	29	29	0	7	20	40	30	10	10
45	74	21	0	5	19	0	100	0	0	1
Total	18	42	27	13	2,239	22	46	23	9	2,956

School	Standard 90851									
	Digital					Paper				
	%N	%A	%M	%E	<i>n</i>	%N	%A	%M	%E	<i>n</i>
1	42	58	0	0	12					0
2	44	44	11	0	9	45	55	0	0	22
3	11	53	29	8	93	25	53	18	5	85
4					0					0
5	4	46	50	0	24	20	60	0	20	5
6	45	41	12	2	58	33	67	0	0	6
7	14	76	10	0	21	13	53	27	7	147
8	2	41	42	15	195	6	35	38	21	190
9	0	18	45	38	56	6	26	40	28	87
10	14	54	25	7	284	18	58	20	4	45
11	27	45	18	9	55	25	33	33	8	12
12	14	71	14	0	21	22	51	20	6	49
13					0					0
14	11	44	38	7	45	0	100	0	0	1
15	0	25	75	0	4	14	45	34	7	148
16	15	63	22	0	27	22	52	16	10	91
17	0	60	35	5	20	19	55	18	8	119
18	9	43	34	14	58	21	59	16	4	264
19					0					0
20					0					0
21	0	50	38	13	8	15	64	16	5	55
22	67	33	0	0	3					0
23					0	28	46	21	5	137
24	8	43	32	17	63	13	46	24	18	85
25	6	42	33	18	33	21	44	23	13	39
26					0					0
27	33	67	0	0	3					0
28	4	61	28	7	54	0	100	0	0	1
29					0	4	41	34	21	80
30	20	62	15	3	71	0	75	25	0	4
31					0	33	50	17	0	6
32					0					0
33	50	50	0	0	2	16	51	33	0	57
34	2	26	40	32	171	0	22	56	22	27
35	9	55	18	18	11	24	54	23	0	71
36	15	52	22	11	126	13	50	31	6	16
37	10	54	23	13	52	18	46	25	11	103
38	5	48	38	10	21	7	48	30	15	46
39	16	72	9	3	67	18	52	22	8	60
40	3	49	44	4	72	50	50	0	0	4
41	100	0	0	0	2					0
42					0					0
43					0	100	0	0	0	1
44					0	10	50	30	10	10
45	33	56	11	0	18	38	54	8	0	24

Total	11	48	29	12	1,759	17	48	25	10	2,097
-------	----	----	----	----	--------------	----	----	----	----	--------------