The moderation of coursework and controlled assessment: A summary

Tim Gill Research Division

Introduction

To ensure consistency and accuracy of marking, awarding bodies carry out moderation of GCSE and A level internally assessed work (e.g., coursework or controlled assessment). Training and instructions are provided by the awarding body to the internal assessors in each centre, including training in task-setting, marking and internal standardisation. Internal standardisation is necessary to ensure the standard is the same across all assessors within a centre.

Awarding bodies are required to modify centres' marks where necessary to bring judgements into line with the required standard. Samples are taken of (internally standardised) candidates' work, across all units and adequately covering the range of attainment within a centre. A moderator re-marks the sampled work, and if there is a difference between the centre's and moderator's marks that is larger than a certain amount (known as the tolerance level) then marks should be adjusted. Should it be necessary to adjust a centre's marks then the magnitude of the adjustments is determined by a regression analysis, based on the relationship between the marks given by the centre and those of the moderator in the sample.

This article summarises the processes undertaken by the Oxford, Cambridge and RSA (OCR) exam board to moderate and, if necessary, adjust the marks of centre-marked coursework and controlled assessments. Some brief data analysis is also presented to give an idea of the extent of moderation and how much difference it makes to candidates' marks.

Moderation and scaling processes

Broad guidelines for the moderation process are set out in the Ofqual Code of Practice document (Ofqual, 2011). More detailed principles and practices were drawn up by the exam boards, as described in an OCR document which provided guidance to centres (OCR, 2010). However, the processes described here refer to those undertaken by OCR only. Other boards may have different processes, so long as they comply with the Code of Practice and the board agreement.

Sampling and moderation

The Ofqual guidelines for sampling student work are quite broad, only requiring that exam boards request samples of work from centres which adequately represent the range of attainment within the centre, requesting additional samples if necessary. They do not specify how this should be done. The OCR procedures are much more detailed, as follows; for each centre taking a coursework unit a sample of (internally marked) scripts (chosen by OCR) are sent to the moderator ('Stage 1'). This sample is drawn from across the range of marks in the centre, and includes the lowest and highest centre marks. A first sample is moderated and if there are no differences above tolerance then no more moderation is necessary and the centre's marks are accepted. However, if one or more differences exceed tolerance then a further sample is moderated ('Stage 2'). If, after this second moderation, the pattern of changes suggested by the moderator is relatively consistent (i.e., it retains the rank order of candidates) then the centre's marks are scaled (see later description). If they are not consistent then it is possible to take a third sample for moderation ('Stage 3'). If after this a valid scaling is still not possible then further options include the moderator re-assessing all candidates in the centre and applying the moderated marks, or the centre re-assessing all work and a new sample being taken.

The size of the sample(s) described above depends on the number of candidates in the centre taking the coursework unit, as shown in Table 1.

Table 1: Sample sizes for different centre sizes

No. of candidates in centre	*Stage 1 (sub) sample*	*Stage 2 (full) sample*	*Stage 3 sample*
1–5	All	All	All
6–10	5	All	All
11–15	6	10	All
16–100	6	10	15
101–200	6	15	20
201+	6	20	25

Scaling

Following the moderation, the scaling adjustments that will be applied are determined through the application of a regression algorithm. The use of regression to determine adjustments is not required by the Ofqual guidelines, and in the inter-board agreement it is only given as an example of how 'automatic' adjustments could be applied. The purpose of the regression algorithm is to determine whether to adjust a centre's marks and if so, by how much. These adjustments will be applied to all candidates in the centre, not just the sample. Only centres where the result of the moderation of the sample was at least one script outside of tolerance go forward to the regression algorithm. Even then it is not certain that it will be necessary to adjust the marks in the centre. If the adjustments suggested by the algorithm are within the tolerance for the unit then the centre marks are accepted. That is, if the adjustment that would be performed would only alter marks by an amount less than or equal to tolerance, then the original marks are close enough to be accepted.

In order to decide how much to adjust a centre's marks by, a regression equation is used to model the relationship between centre and

Table 2: Example application of moderation and scaling procedure

Centre mark (X)	**36**	**36**	**34**	**33**	**31**	**29**	**27**	**24**	**21**	**18**	**16**	**16**	**11**	**7**	**6**
Moderator mark (Y)	34	32	32	31	30	29	28	23	22	21	17	16	14	8	8
Mod – Centre (Y-X)	-2	-4	-2	-2	-1	0	1	-1	1	3	1	0	3	1	2
Predicted mark (Ŷ)	33.9	33.9	32.2	31.3	29.7	28.0	26.3	23.8	21.3	18.7	17.1	17.1	12.9	9.5	8.7
Regression mark	34	34	32	31	30	28	26	24	21	19	17	17	13	10	9
Regression – Centre	-2	-2	-2	-2	-1	-1	-1	0	0	1	1	1	2	3	3
Final mark	**34**	**34**	**32**	**31**	**30**	**28**	**26**	**24**	**21**	**19**	**17**	**17**	**13**	**10**	**9**

moderator mark[1]. The form of this equation used by the algorithm is as follows:

$$Y = aX + b$$

Where Y is the moderator mark, X is the centre mark and 'a' and 'b' are the regression parameters. For each centre mark (X) in the sample a predicted adjusted mark (Ŷ, also known as the 'regression mark') is generated from this equation. The 'a' and 'b' parameters are set so as to minimise the average of the squared difference between each moderator mark and predicted mark in the sample.

The magnitude of the adjustment (if it is deemed necessary) is the difference between the centre's mark and the regression mark. Often, the regression mark is not a whole number, in which case it is rounded up or down. Take the following example, for a unit of maximum mark of 40 marks, with a tolerance of 2 marks and the following regression equation:

Moderator mark = 0.84*Centre mark + 3.62.

Table 2 presents the centre and moderator marks for the sample and follows through the procedure to get the final marks applied to these marks.

The regression equation generates a predicted mark as shown in Table 2 (Ŷ, to 1 decimal point). This is rounded up or down to generate the regression mark. This becomes the final mark if the algorithm determines that an adjustment to the centre's marks is necessary. A final check is made of whether any of the adjustments are outside of tolerance. If none are, then the centre marks are accepted. In the example in Table 2 there were two candidates with proposed adjustments of 3 marks, greater than the tolerance of 2 marks (highlighted by the red squares), so the decision would be to adjust the centre's marks.

This example is displayed graphically in Figure 1.

The crosses represent the centre and moderator mark for each candidate. The blue line is the regression line, indicating the proposed changes to centre marks. This is not straight because of the rounding up or down that is necessary to enable the mark adjustments to be whole numbers. Note that the regression line tapers at the bottom of the mark range so that candidates with a mark of zero have their mark unadjusted.

Finally, the straight lines are bands for the level of tolerance for this unit. These bands are two marks either side of the 'identity' line (not shown, but where the centre and moderator marks are equal). This figure shows that in the sample there were three scripts with marks outside of tolerance (outside of the bands). At the bottom of the mark range there were two scripts for which the regression line suggests an adjustment that would be greater than tolerance. This means that this centre's marks would need to be adjusted.

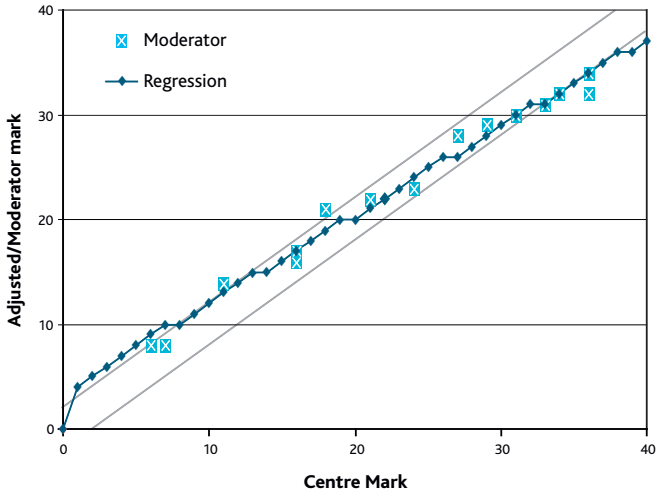


Figure 1: Plot of centre and moderator marks with regression marks, and tolerance bands

Criteria for automatic scaling

Once the algorithm has determined that adjustments are necessary, these are applied to all candidates in a centre automatically, as long as some specific criteria are met. These criteria are not required by Ofqual regulations but were generated by OCR to ensure some checks are carried out on the scaling undertaken. The aim of the criteria is to flag up any scaling decisions that are particularly out of the ordinary in some way (e.g., large adjustments to marks), or might be unfair to some candidates. If at least one of these criteria is not met, the centre is flagged up so that OCR Operations staff can look in more detail at the proposed scaling decision and decide whether or not it is valid. The criteria are:

1. No 'unusual marks' in the sample. Unusual marks are those where the difference between the regressed mark and moderator mark is larger than 10 per cent of the maximum mark.
2. The average of the squared difference between the moderator marks and the regression marks is less than or equal to 3.5. This is so that centres where the adjustments to candidates are very different to those suggested by the moderator are not included automatically.

1. For marks between 10% and 100% of the maximum mark a simple linear regression is used. For the bottom 10% the regression line is a curve so that the centre and moderator marks converge at zero.

3. No large differences between centre and moderator marks at the extremes of the sample. A large difference at the extremes might mean that excluding this candidate would have a big impact on the scaling decision and adjustments.
4. More than one mark outside of tolerance. This is because if only one mark was greater than tolerance then excluding this candidate (which is an option open to Operations staff when reviewing the recommended mark adjustments) would change the scaling decision from adjusting to not adjusting.
5. The average absolute adjustment applied to all candidates in the centre is not greater than 15 per cent of the maximum raw mark. This is to ensure that any particularly large adjustments are flagged up.
6. Correlation between centre and moderator marks is at least 0.75. A correlation lower than this would suggest a valid scaling would be difficult.

If it is decided that the proposed scaling is not valid then there are two options available. First, where there are unusual marks in the sample, it is possible to exclude these candidates and re-run the regression to see what the impact is on the proposed adjustments. If a candidate is having a detrimental effect on the adjustments for all other candidates then it might be justified to exclude them. However, candidates should not normally be excluded if it would change the scaling decision from applying to not applying adjustments.

If it is still not possible to create a valid scaling outcome using the regression algorithm then the procedures allow for manual scaling. This means manual adjustments are made at each mark point without recourse to the regression algorithm.

Once the scaling has been determined it is applied to all candidates in the centre, not just those in the sample. This is communicated to the centre in the form of a Banding Report, showing the scaling that needs to be applied to different bands of centre marks. The report covers the whole of the mark range, whether or not the centre has any candidates with a mark in a particular band. An example is shown in Table 3.

Table 3: Example Banding Report

Marks From–To	*Scaling Factor*
34–40	-2
28–33	-1
21–27	0
14–20	1
8–13	2
1–7	3

Data analysis

This section explores some background data in relation to moderation of coursework by OCR. The data comes from the June 2012 session.

Extent of moderation

Several of the qualifications offered by OCR involve some components that require moderation. Table 4 presents the number of components in each qualification that were moderated in the June 2012 session.

Table 4: Number of moderated components by qualification, June 2012

Qualification	*Components*
A level	126
GCSE	95
Principal learning (Level 1, 2 and 3)	74
Entry level certificate	18
Other	6
All	**319**

Table 5: Summary of moderated components, June 2012

Moderated components	319
Moderated centres	35,011
Regressed centres (%)	28.0
Scaled centres (%)	22.5
Candidates in scaled centres	188,091
Scaled candidates	167,763

Table 5 presents the total number of components, centres and candidates that were affected by scaling in June 2012. It also presents the percentage of centres taking units subject to moderation whose marks were scaled.

Thus, there were 319 components that were moderated and 35,011 centres subject to moderation. Of these centres, 28% were found to have at least one difference between centre and moderator mark that was larger than the allowed tolerance, meaning the marks went through the regression algorithm. However, only 22.5% of moderated centres actually had scaling applied. The reason for this difference is that some of the regressed centres had all the 'regressed' marks inside of tolerance (see earlier description). The 'candidates in scaled centres' figure in the table includes candidates whose mark did not in fact change, because their scaling adjustment was 0 marks. The 'scaled candidates' figure only includes students whose marks were adjusted.

An analysis was also undertaken of the percentage of centres in each component that were regressed. The results showed that there were 35 components where no centres were regressed (i.e., all centres marks accepted) and 18 components where all centres were regressed. Most of these components were taken by a very low number of centres (fewer than 10), but there was one component with 89 centres, none of which were regressed. There was also a component with just one centre regressed out of 191 (0.5%). Excluding components where all centres were regressed (each of which consisted of fewer than five centres), the highest percentage of centres regressed was 80.3 (196 out of 244). Figure 2 presents the distribution of the percentage regressed, for components with at least 50 centres.

In terms of the percentage of centres within each component that were actually scaled, there were 42 components where none of the centres were scaled, one of which had 191 centres and one 89 centres. Otherwise the numbers of centres for these components were generally very low. There were 13 components where all centres were scaled, but these were all components with very few centres. Of the components with more than 50 centres, the highest percentage of scaled centres was 76.7% (69 out of 90). The lowest percentage scaled was 1.2% (4 out of 343).

Figure 3 presents the distribution of the percentage scaled, for components with 50 or more centres.

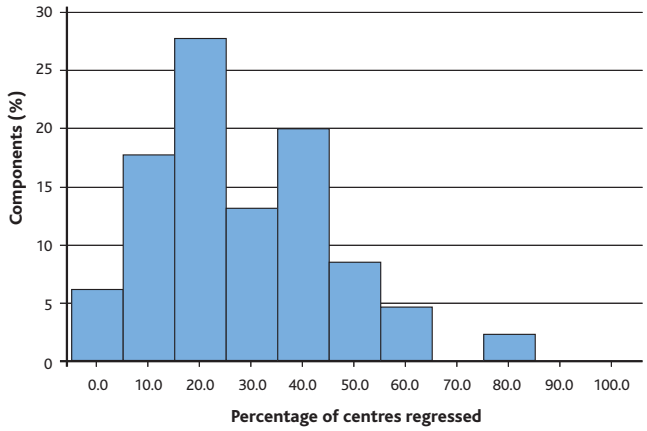


Figure 2: Distribution of the percentage of centres regressed (components with 50 or more centres)

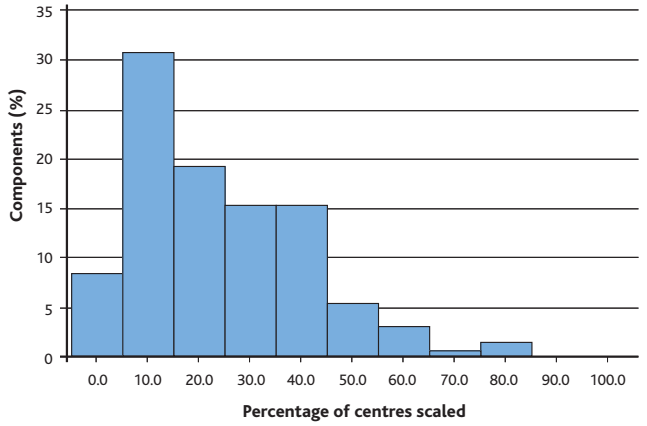
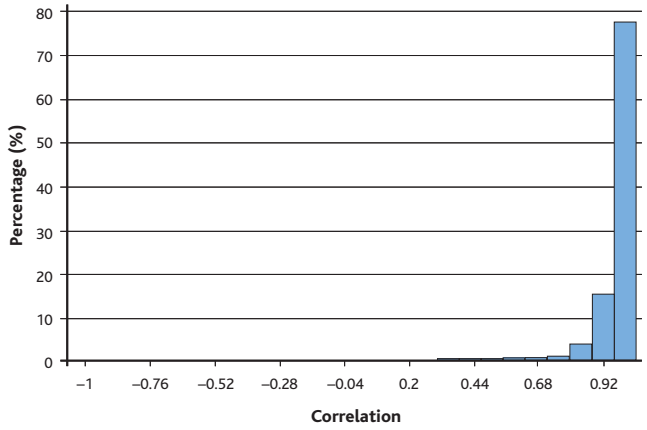


Figure 3: Distribution of the percentage of centres scaled (components with 50 or more centres)



Correlation	*Centres*	*Percentage of centres*
<0.75	135	2.0
0.75–0.80	60	0.9
0.80–0.85	119	1.8
0.85–0.90	252	3.8
0.90–0.95	656	9.8
>0.95	5,459	81.7
All	6,681	100.0

Figure 4: Distribution of correlation coefficients in scaled centres

Correlation between centre and moderator marks

One way of assessing the level of agreement (in terms of rank order of candidates) between centre and moderator marks is through a correlation coefficient. This was calculated for each (scaled) centre in each component in the June 2012 data. These correlations used the marks for the sampled scripts only, as these were the only scripts with a moderator mark. Figure 4 presents the distribution of correlation coefficients.

Almost 82% of centres had a correlation of greater than 0.95 and 91.5% had a correlation of greater than 0.90. Thus, in terms of the rank order of candidates within a centre, there is usually a lot of agreement between centre and moderator mark even in the centres which were scaled. However, this doesn't necessarily mean that there is a high level of agreement over the marks. It may be that the centre marks tend to be consistently higher than moderator marks. This explains why a substantial percentage of centres are scaled, even though correlations tend to be very high.

Adjustments to marks

Another important aspect of the scaling process is how large the adjustments to candidates' marks are. Table 6 summarises the changes to candidates' marks as a result of scaling, (a negative figure means a reduction in the mark given to the candidate). This includes all candidates in centres that were scaled, not just those in the sample.

Overall, adjustments were much more likely to be negative than positive, with just 5.8% of adjustments greater than 0, compared with 83.4% less than 0. The remaining 10.8% of candidates had no adjustments to their marks, despite being in centres where some adjustments were necessary.

Table 6: Summary of adjustments made to candidates' marks

N	*Mean*	*SD*	*Min*	*Max*
188,091	-3.9	3.9	-60	40

This analysis was repeated for the adjustment as a percentage of the maximum mark for the component. Figure 5 presents the distribution of adjustments. The mean adjustment was -6.7% with a minimum adjustment of -65% and a maximum of 58.3%.

A further analysis was undertaken of the mean adjustment to marks for each individual component. There was a fairly wide range of adjustments, with the biggest negative adjustment on average for a component being -21.6 marks (although this component was only taken

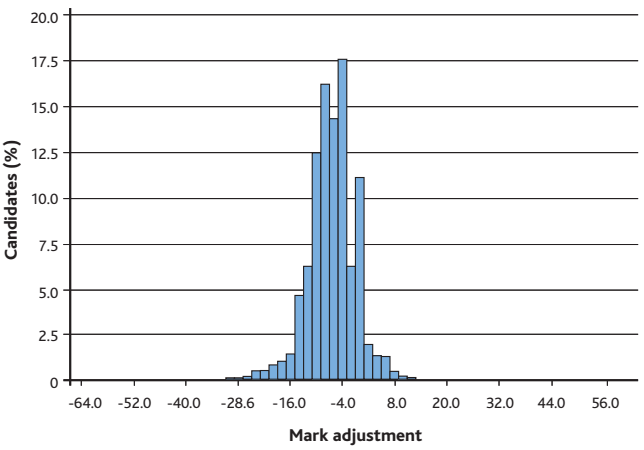


Figure 5: Distribution of mark adjustments (percentage of maximum mark)

by 19 candidates), whilst the most positive was +8.2 marks. The biggest mean adjustment for a component with more than 100 candidates was -10.8 marks. However, this component had a maximum mark of 120, so the average adjustment was less than 10%.

Figure 6 presents the distribution of mean adjustments for each component in terms of percentage of maximum mark (restricting to components with more than 100 candidates). The largest negative adjustment was -18.7% (-9.3 marks for a component with a maximum mark of 50). The largest positive adjustment was 8.2% (+8.2 marks for a component with a maximum mark of 100).

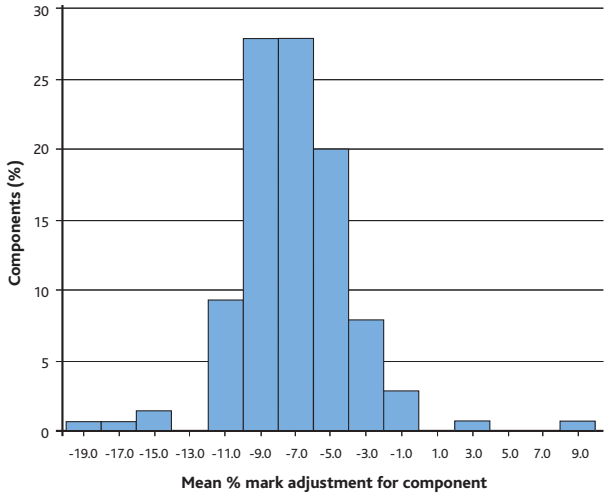


Figure 6: Distribution of mean adjustment to mark (as a percentage of maximum mark) by component

Extent of automatic scaling

As previously noted, in order for the scaling to proceed automatically without being checked, a number of criteria need to be met. For the June 2012 session the number of centres where at least one of the criteria was failed and the scaling outcome was checked was 2,102 (31.3% of all centres scaled). Table 7 presents the number of centres failing each criterion. The criteria are not mutually exclusive, so it is possible for centres to fail more than one.

Table 7: Frequency of centres where criteria for automatic processing not met

Criterion	*Count*	*Percent*
1	108	1.61
2	980	14.58
3	403	5.99
4	482	7.17
5	676	10.06
6	139	2.07

Of the centres that were checked, 14.6% (306 centres) had their scaling adjusted manually (either by excluding candidate(s) from the regression and re-running or by deciding on the scaling to be applied at each mark point without recourse to the regression algorithm). This is 4.6% of all centres that were scaled.

Table 8 shows the frequency of the centres where a given number of criteria were not met. For instance, the first row shows that all the six criteria were met in 68.73% of centres. Around 23% of centres failed to meet just one criterion and relatively few centres (7.75%) failed on two or more of the criteria. The final row in the table indicates that 2,102 centres failed at least one criterion.

Table 8: Frequency of all criteria not met, in centres that were scaled

Count of criteria	*Count*	*Percent*	*Cumulative count*	*Cumulative Percent*
0	4,621	68.73	4,621	68.73
1	1,581	23.52	6,202	92.25
2	387	5.76	6,589	98.01
3	106	1.58	6,695	99.58
4	25	0.37	6,720	99.96
5	3	0.04	6,723	100.00
At least 1	2,102	31.27	-	-

Discussion

This article has outlined the purpose and processes involved in the moderation of coursework and controlled assessment at OCR. It has also demonstrated the extent of moderation undertaken, both in terms of the percentage of centres moderated and the levels of adjustments implemented.

It is worth noting that moderation is not meant to be the same as re-marking of work. It would not be possible for all the work in a centre to be re-marked because of the number of candidates taking these units. Instead, as described earlier, moderators re-mark a sample of the work, and use the relationship between moderator mark and centre mark in the sample to estimate what adjustments should be made to candidates' work in the whole centre. This means that some candidates whose work had been moderated will end up with a mark that is different from the mark they 'should' have received (as given by the moderator). However, the principle here is to be as fair as possible to all candidates in the centre, including those whose work has not been moderated. As we don't know the actual mark these candidates should have received, the best estimate is that generated by the relationship between moderator and centre mark (as long as that relationship is reasonably consistent across the mark range).

This article has shown that most centre marking is of the required standard: less than one quarter (22.5%) of centres taking moderated components needed to have their marks adjusted. Furthermore, when adjustments were necessary, these tended to be small (although there were some exceptions) and the correlations between centre and moderator mark (within a centre) were mostly very high. This suggests that the guidelines and training given to assessors within centres by OCR (in terms of marking and internal standardisation) are generally clear and understandable. We have also shown that only around 1 in 7 (14.6%) scaling decisions that were flagged as requiring checking were subsequently changed. This suggests that, on the whole, the regression algorithm works well in generating fair adjustments to candidates' marks.

However, it is also worth noting that it was much more common for centres to be generous than severe in their marking, in comparison to the moderator mark. This is perhaps not surprising, as teachers want their students to do as well as possible in their qualifications.

Finally, two further points about how OCR ensures that moderation is as fair and accurate as possible are worth mentioning. First, Ofqual regulations require that moderators must be trained and undertake

standardisation and have their moderation standards checked by a senior moderator. Those judged to be unsatisfactory will no longer be allowed to undertake moderation and candidates' work in centres that they moderated will need to be re-moderated. Secondly, if a centre has its candidates' work scaled and is unhappy with the adjustments made, they can request a review of the moderation (for a fee). If it is determined that the original moderation is not acceptable then a revised moderation is implemented instead.

References

Ofqual (2011), *GCSE, GCE, Principal Learning and Project Code of Practice: May 2011*. Coventry: The Office of Qualifications and Examinations Regulation. Retrieved from http://ofqual.gov.uk/documents/gcse-gce-principal-learning-and-project-code-of-practice/

OCR (2010). *Moderation of GCE, GCSE and FSMQ centre-assessed units/components: Common principles and practices*. Cambridge: Oxford, Cambridge and RSA.

Reflections on a framework for validation – Five years on

Stuart Shaw Cambridge International Examinations and **Victoria Crisp** Research Division

Abstract

In essence, validation is simple. The basic questions which underlie any validation exercise are: what is being claimed about the test, and are the claims warranted (given all of the evidence). What could be more straightforward? Unfortunately, despite a century of theorising validity, it is still quite unclear exactly how much and what kind of evidence or analysis is required in order to establish a claim to validity. Despite Kane's attempts to simplify validation by developing a methodology to support validation practice, one which is grounded in argumentation (e.g., Kane, 1992), and the "simple, accessible direction for practitioners" (Goldstein & Behuniak, 2011, p.36) provided by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014), good validation studies still prove surprisingly challenging to implement.

In response, a framework for evidencing assessment validity in large-scale, high-stakes examinations and a set of methods for gathering validity evidence was developed in 2008/2009. The framework includes a number of validation questions to be answered by the collection of appropriate evidence and by related analyses. Both framework and methods were piloted and refined. Systematic implementation of the validation framework followed which employs two parallel validation strategies:

1. an experimental validation strategy which entails full post-hoc validation studies undertaken solely by research staff
2. an operational validation strategy which entails the gathering and synthesis of validation evidence currently generated routinely within operational processes.

Five years on, a number of issues have emerged which prompted a review of the validation framework and several conceptual and textual changes to the language of the framework. These changes strengthen the theoretical structure underpinning the framework.

This paper presents the revised framework, and reflects on the original scope of the framework and how this has changed. We also consider the suitability and meaningfulness of the language employed by the framework.

Validation: a task too far?

Samuel Messick's extended account of validity and validation came to dominate the educational and psychological measurement and assessment landscape of the 1980s and 1990s. Instigated by Loevinger (1957), developed and articulated by Messick (1989), and endorsed through the support of significant allies including Robert Guion, Mary Tenopyr and Harold Gulliksen, the essence of validity came to be understood as being fundamentally a unitary concept. Messick's landmark treatise on validity published in the textbook *Educational Measurement* (Messick, 1989) represented the culmination and enunciation of a paradigm shift towards a unified view of validity as articulated in the description of modern construct validity. Measurement was to assume centre stage and came to be the foundation for all construct validity. Since that time, mainstream scholars have consistently affirmed the 'consensus' concerning the nature of validity (e.g., Shepard, 1993; Moss, 1995; Kane, 2001; Downing, 2003; Sireci, 2009) described in the maxim: all validity is construct validity. If validity pivots upon score meaning then by extension construct validation, that is, scientific inquiry into score meaning, is to be understood as the foundation for all validation inquiry. Hence, "... all validation is construct validation." (Cronbach, 1984, p.126).

Tests were to be evaluated holistically, on the basis of a scientific evaluation into score meaning. This approach was to have profound implications for all validation effort. Messick (1998, pp.70–71) seemed to imply that every kind of validation evidence is not only *relevant* but also *necessary* for every validation. Construct validation was to entail scientific theory-testing premised on multiple evidential sources. If the scope of modern validity theory was to be enlarged in an attempt to embrace a full evaluative treatment of consequences (as many, though not all, leading theorists of the day argued and continue to argue) then validation would require monumental effort especially if it was to include an exploration of unintended consequences.

The argument-based approach to validation – as championed by Kane (e.g., 1992, 2001, 2004, 2006, 2013), was an attempt to simplify both validity theory and validation practice. Recognising the difficulties in translating construct validity theory into construct validation practice, Kane rejects the idea that all kinds of evidence are required for every