Standardisation methods, mark schemes, and their impact on marking reliability

February 2014

Alpha*Plus* Consultancy Ltd

Ofqual/14/5380

This report has been commissioned by the Office of Qualifications and Examination Regulation.

Contents

List of tables

1 Executive summary

This document is the report of a literature review carried out by Alpha*Plus* Consultancy Ltd. for Ofqual in the summer and autumn of 2013. The review looked at three areas in order to understand possible factors that affect marking reliability: methods of carrying out standardisation of marking, features of mark schemes and, as background, insights from wider research concerning online meetings and web-based training.

Repeated searches of diverse sources such as: journals, websites, internet databases and so on turned up 115 articles that had the potential to be included in the study, of which 76 were found to be particularly relevant and were therefore analysed in more detail.

UK high-stakes assessment practice, as exemplified by the GCSE and A level Code of Practice, indicates that standardisation is a multi-step process with a range of detailed prescriptions on awarding organisations (AOs). However, there are relatively few prescriptions in respect of the mode of standardisation; remote/e-facilitated standardisation is neither endorsed nor prohibited.

Several UK AOs have moved some of their marker standardisation to an online mode of delivery. This follows the widespread move to online marking.

In the standardisation strand, several reviewed studies purport to show that moving marker training[1] online, and/or remote does not have a deleterious effect on marking accuracy. The studies also show major gains in logistical terms (quicker, cheaper training and more marking per unit time – which presumably also makes it cheaper). However, as with 'mainstream' marker training research, the evidence on particular features of training that are associated with gains or losses of marking accuracy is neither coherent nor strong.

There is not a clear pattern in respect of markers' perceptions of online standardisation; some like it, others do not. In some sets of findings there was a dis-association between perception of the innovation, and its actual impact. In at least one study, markers who benefited from training didn't like it, whereas those whose marking was not improved by the training did. We also note that early experiences of on-line marking and standardisation may give little indication of how such methods would impact on markers once established over a longer period of time.

The observation that a community of marker practice might not be in causal association with marking reliability is discussed. It is suggested that, in fact, such a causal link might not be the most important justification for maintaining a community of practice[2]. Rather, it might be that maintaining teachers' engagement with marking, and hence with the examinations system, is a better justification for retaining a 'social' aspect to marker standardisation.

A range of statistical techniques is employed to study the effects of different standardisation methods on marking accuracy. Many studies use classical methods, which can be extremely useful even if they are inherently limited. Other techniques bring different benefits, although some bring disadvantages as well. Results from some models, for example, can appear 'bitty' in some studies.

There is relatively little detailed research into mark schemes and their effect on the reliability of marking, and still less in which there are clear conclusions to be drawn. However, it is possible to draw out some particular ideas. First it has been suggested that the mark scheme (or at least a prototype version of it) should precede the construction of the item. Moreover, the assessment design process can be seen as starting with the identification of the evidence for the levels of

[1] Although the review title is 'standardisation', the majority of the results were returned against the search terms 'marker' or 'rater training'.
[2] Or 'a shared understanding between professionals'.

required performance followed by the construction of a suitable rubric (mark scheme) to formalise the levels identified, before the devising of items and tasks.

Perhaps the single most consistent underlying factor identified in all the work that relates to the effect of mark schemes on reliability is the need for clarity and avoiding unnecessary complications. This applies whether the mark scheme in question is for an objective item with a single correct answer or a levels-based scheme for an extended piece of writing (or an artefact). It is, however, important to realise that the pursuit of simplicity should not involve a threat to validity, a point made in several of the relevant papers. Some authors argue for a clear statement of the principle behind the award of credit rather than attempting to anticipate the entire outcome space[3].

It has been reported that the mark schemes for key stage 3 English assessment provided four different sources of information to help the marker make decisions: the assessment objectives a question is meant to be testing, illustrative content, performance criteria (essentially levels-based descriptions) and exemplar responses. The author noted evidence that practice focused principally on the performance criteria, occasionally checked against the exemplar materials, and also reported that despite – or because of – all the information, markers remained unclear on a number of key issues. It seems that, although the mark scheme did provide a statement of the key principle against which each question was to be assessed (the Assessment Objectives) this was obscured by the quantity and variety of detail provided.

This idea also applies to levels-based mark schemes. Whether or not a holistic or analytic approach is preferred (and the evidence is unclear as to which is more effective in achieving reliable marking) the key is to minimise the cognitive demand on the markers. The pursuit of clarity about what is required is important in helping to avoid the use of construct irrelevant factors when arriving at an assessment decision. It has been noted that assessors often make difficult choices between two levels on a rubric scale by using extraneous factors. It is clearly preferable to give every assistance in using relevant factors. However, the temptation to achieve this by devising highly detailed mark schemes should be resisted.

[3] Outcome space relates to the range of responses from poor to good responses that students will produce in response to an assessment task. The more accurately an assessment designer anticipates the range of responses a body of students will produce, the more valid the assessment task.

2 Introduction

In 2012 Ofqual committed to carry out a programme of work looking into the quality of marking in general qualifications in England (Ofqual, 2012). The aims of this work are:

- To improve public understanding of how marking works and its limitations
- To identify where current arrangements work well (and where they don't)
- To identify and recommend improvements where they might be necessary (*ibid.*)

The quality of marking review focusses on general qualifications (GCSEs, IGCSEs, A levels, International A levels, International Baccalaureate Diploma and the Pre-U Diploma). In July 2014 Ofqual commissioned Alpha*Plus* Consultancy Ltd. to conduct a literature review on the impact that different standardisation methods have on marking reliability and marker engagement and the features of mark schemes which are most associated with accuracy and reliability of marking. Ofqual also required the review to contain a brief study of online training and its (potential) impact on standardisation.

3 Findings

3.1 Standardisation

3.1.1 Impact of different forms of standardisation on quality of marking

In this section, we summarise findings from the relatively small number of studies that we consider to be well designed and to provide robust evidence in respect of the effect of different marker training/standardisation methods.

We report findings concerning standardisation methods and marker engagement separately, because engagement and effectiveness are not always related in a straightforward manner; for example, there are training programmes that recipients appear to like, but which apparently deliver little or no improvement in marking quality, as well as the converse situation.

Wolfe, Matthews and Vickers (2010)[4] designed a research exercise using secondary school students' essays which were composed in response to a state-wide writing assessment in the USA (*ibid.*, at p. 6). Their study compared marker performance amongst three conditions: distributed online, regional online and regional face-to-face training[5]. These conditions were defined as follows:

> (a) rater training that is conducted online followed by scoring that occurs through a computer interface at remote locations (referred to here as an *online distributed training* context),
>
> (b) rater training that is conducted online followed by scoring that occurs through a computer interface, both of which take place at a regional scoring center (referred to here as an *online regional training* context), and
>
> (c) face-to-face training followed by scoring that occurs through a computer interface, both of which take place in a regional scoring center (referred to here as a *stand-up regional* context). (*ibid.*, at p. 5)

They found that, on their defined score-quality indices, the online distributed group assigned ratings of slightly higher quality in comparison to the ratings assigned by the two other groups. However, such differences were not statistically significant (*ibid.* at p. 13).

Whilst there were not significant differences between the quality of marking in the three modes, there was a clear difference in respect of the time that the face-to-face training took. In general, this mode took three times longer than either form of online training. This difference was statistically significant and the effect size was large when judged against established guidelines (*ibid.* at p. 14).

Chamberlain & Taylor (2010) measured and compared the effects of face-to-face and online standardisation training on examiners' quality of marking, in a research study, utilising history GCSE scripts. They found that both face-to-face and online training had beneficial effects but that there was not a significant difference between the modes. Indeed, they suggested that improvements were quite modest, and posited a 'ceiling effect' in that markers were already marking with high quality, and thus there was not much room for improvement (*ibid.* at p. 7).

Knoch, Read and von Randow (2007) compared the effectiveness of the face-to-face and online methods for re-training markers on the writing assessment programme at a New Zealand University. Once again, both training modes brought markers closer together in their marking. There was some indication that online training was slightly more successful at engendering marker consistency. In

[4] See also: Wolfe and McVay (2010).
[5] They also refer to the last condition as 'stand-up training'.

contrast, face-to-face training appeared to be somewhat more effective at reducing marker bias (*ibid.*, at p. 41).

Elder et al (2007) also studied the impact of an online marker training programme in a New Zealand university; in this case, the work was based on a Diagnostic English Language Needs Assessment (DELNA). They stated (with seeming regret) that: 'the effort involved in setting up the program did not pay off' (*ibid.*, at p. 55). Although somewhat reduced following training, severity differences between markers remained significant, and negligible changes in marker consistency were achieved.

Way, Vickers and Nichols' (2008) conference paper commented upon previous research, such as that of Vickers and Nichols (2005). Vickers and Nichols (2005)'s study of American seventh-graders' written responses to a reading item found that the online and face-to-face trained groups were able to provide marking of similar quality, but that those trained online were able to mark about 10 per cent more responses than the face-to-face trained group in the same time period (Way, Vickers & Nichols, 2008, pp. 6 – 7).

Knoch (2011) studied marking in a large-scale English for Specific Purposes (ESP) assessment for the health professions over several administrations. Data were available on eight sittings of the ESP assessment, with training conducted via phone or email, or – in the final training session – by email, or interview. This longitudinal approach[6] is unusual in the context of studies considered here; more longitudinal information could potentially tell us whether effects are long-lasting – reducing the effect of markers' existing expertise, which may endure in simulated intervention studies. The downside of Knoch's (2011) study, for those seeking to understand the impact of online standardisation, is that she had no face-to-face/conventional condition to control against the electronically-mediated training.

The feedback gave information adapted from the FACETS Rasch model analysis software. As its name suggests, that software models measurement inaccuracy in respect of different facets. In terms of severity, bias and consistency, the training was found to deliver no more benefit than random variance. This was true of speaking and writing markers equally (*ibid.*, at p. 196).

Xi and Mollaum (2011) report more success than Knoch (2011) with their training programme. They investigated the scoring of the Speaking section of the Test of English as a Foreign Language Internet-based test by markers who spoke English and one or more Indian languages. Their study contained treatment and control groups (the 'regular' and special' training groups) (*ibid.*, at p. 1232). The training was realised via a special downloadable computer program designed to replicate operational procedures (*ibid.*, at p. 1233). To that extent the study showed that computerised training could be effective. However, the distinction between the two groups was in terms of the composition of exemplar speaking samples; in the special group more prominence was given to native speakers of Indian languages.

The study did show the effectiveness of the special training procedure, with marking quality being significantly improved in the special training approach. However, this demonstrated the effectiveness of including increased numbers of Indian language native speakers in the standardisation sample of speech, rather than demonstrating the effectiveness of online training per se.

In their Iranian university, English as a Foreign Language context, Fahim and Bijani (2011) developed a training package that was provided to markers on CD-ROM for them to work on at home. Fahim and Bijani evaluated a novel (for them) training implementation's potential to standardise markers' severity and to reduce individual biases. In fact, the study showed that the training was able to

[6] Knoch calls it to a 'time series design' (2011, p. 187).

move markers to a more appropriate standard and to reduce bias, although not to eliminate it entirely (*ibid.,* at p. 11).

3.1.2 Diversity of studied standardisation training initiatives

In the previous section we have summarised the outcomes of the relevant studies that our searches have thrown up. However, even though it is valid to report those outcomes, it is important to acknowledge the sheer diversity of contexts and approaches taken, the very different examinations studied and hence the challenges that meet any attempt to generalise from findings.

Wolfe, Matthews and Vickers (2010, p. 8) had markers marking to a four-point holistic 'mark scheme' (or 'rubric', to use the term with which they would be more familiar). The face-to-face trainer used standardised annotations on exemplar scripts and then these same annotations were presented electronically during the online training (*ibid.*). At the regional training site, markers could either ask questions face-to-face or online – depending upon the experimental group to which they belonged. In contrast, the 'distributed' markers' questions would be answered by email or phone.

Chamberlain and Taylor (2010) designed their online application to combine specific electronic enhancements and some pre-existing features of the face-to-face meeting. They felt that this was a more legitimate approach than merely attempting to replicate the functions of face-to-face interaction in the online environment (*ibid.*, at pp. 2 – 3). Knoch, Read and von Randow (2007) went further (perhaps) than Chamberlain and Taylor in that same direction – in that she adapted outputs from the FACETS Rasch analysis software (see below at p. 10)[7]. She used a range of graphic and tabular presentations to inform markers of their severity, internal consistency, central tendencies and halo effects. As such, in interpreting Knoch, Read and von Randow's findings, we might conclude that they give us insight into the use of technology for online training – or equally, they might tell us about what happens when a psychometrician tries to present subject matter experts with statistical information.

There is corresponding diversity in the size of studies. For example, Knoch, Read and von Randow (2007) had 16 markers mark 70 scripts each. Vickers and Nichols (2005), in contrast, had 63 markers mark over 35,000 scripts (Way, Vickers & Nichols, 2008, p. 6).

[7] Knoch (2011) used similar training materials.

Notwithstanding such differences, we present a summary of the findings in Table 1, below.

Study authors	Summary findings	Significant effect	Comment
Wolfe, Matthews and Vickers (2010)	No significant differences between the quality of marking in the three modes.	No	
	Clear difference in respect of the time that the face-to-face training took.	Yes	Large effect size.
Chamberlain and Taylor (2010)	Both face-to-face and online training had beneficial effects	No	May be a ceiling effect; marking good to start with, so hard to improve.
Knoch, Read and von Randow (2007)	Online training slightly more successful for marker consistency	No	
	Face-to-face training appeared to be somewhat more effective at reducing marker bias	No	
Elder et al (2007)	Severity differences between markers remained significant	No	Study sought improvements of online training. Not a control group study.
	Negligible changes in marker consistency	No	
Vickers and Nichols (2005)	Online and face-to-face groups provided marking of similar quality (reliability and validity)	No	
	Online group able to mark 10 per cent more responses than the face-to-face group in same time.	Not stated	
Knoch (2011)	Online training deliver no more benefit than random variance in respect of severity, bias and consistency.	No	Longitudinal study, but no control group
Xi and Mollaum (2011)	Special downloadable computer score program provided benefits.	Yes	Control design concerned L1 of speakers and markers, not online method.
Fahim and Bijani (2011)	Training reduced biased and harshness to great extent, but did not eliminate it.	No sig test in study	

Table 1: Summary of findings on online standardisation

We add this table to extract maximum information from the reviewed studies. Of course, non-significant findings are often considered to amount to no findings at all. However, we retain the table as a source of useful guidance – so long as caveats around the significance of findings are taken on board.

3.1.3 Perceptions of marker training and standardisation

In this section we report on markers' perceptions of the training activities in which they participated. As with the improvements to marking quality reported in the section above, there is no single direction in the results; some innovations seem to find favour with participants, others do not.

For clarity, we report perceptions findings separately from quality of marking findings. This is because the connection between markers' perceptions of training initiatives, and those initiatives' impact on marking quality is something that frequently occupies (indeed troubles) researchers. Quite a few researchers found that there was little, or a counter-intuitive, relationship between marking quality improvements facilitated by a training initiative and markers' perceptions of that same initiative.

This lack of relationship is illustrated most starkly in Knoch (2011). The researcher compared the success of feedback given during marking (for improving marking quality) with respondents' attitudes to it. She delineated four groups of success-attitude combinations:

- Feedback successful, positive about feedback
- Feedback successful, negative about feedback
- Feedback unsuccessful, positive about feedback
- Feedback unsuccessful, negative about feedback

Unfortunately (from Knoch's point of view), around 70 per cent of the markers fell into the middle two rows; that is, those for whom the intervention had no impact were positive about it, or those for whom it worked nevertheless didn't like it[8] (*ibid.*, at pp. 195 – 196).

Wolfe, Matthews and Vickers (2010, p. 15) found that online distributed and face-to-face regional training groups had more positive attitudes to their training mode than the regional online group. However, these differences were not statistically significant. Further, they reported that the face-to-face trainees seemed to ask more questions about content, whereas online trainees made more requests about logistics, user interface, and so on.

In common with other researchers, Knoch, Read and von Randow (2007) could not find a straightforward relationship between training mode and marker perception. However, they used their opinion data to develop this summary of advantages and disadvantages of the respective training modes:

Advantages	Disadvantages
Online marker training	
Motivating	Tiring, requires greater concentration
Can be done in marker's own time and space	Impersonal, no one to compare ratings with, isolated
Marker is able to take breaks at any point	Hard to read off screen/strain on eyes
No interruption by dominating people, does not affect shy people	
Quick, immediate feedback on rating behaviour	
Objective	
Face-to-face marker training	
Interaction with other markers, discussions	Inconvenient, has to be done at certain time
Easier to compare between scripts	Tiring
Fun, sociable	Markers might be influenced by others and change score because of this

Table 2: Perceived advantages and disadvantages of training modes[9]

Knoch, Read and von Randow (2007, p. 42) also surmised that markers' reactions to online training might be a function of various personal dispositions and circumstances, such as: extroversion/introversion, attitudes to computers, availability of free time, etc.

[8] In review, an alternative interpretation has been suggested; 'Is this surprising since those who experience a challenge to their practices won't like it and those who don't detect a challenge to change practice will feel fine?'
[9] Based on Knoch, Read and von Randow (2007, p. 40).

Wolfe, Matthews and Vickers (2010, pp. 6 – 7) pointed out that, in addition to markers' perceptions of training initiatives, other stakeholders' perceptions needed to be taken into account. They stated that it was easier to get a client to 'sign off' online training materials than it was to get that same client to observe one or more face-to-face training meetings. They also ruminated that, in large-scale testing, it was logistically difficult to offer markers free choice between face-to-face and online standardisation training. Rather, it made more sense for each mode to find a way of delivering the effective elements of its sibling (*ibid*.).

Fahim and Bijani (2011, pp. 11 - 12) did not find a clear direction in their findings between marker quality and perception. They had some evidence (although not a significant finding) that those with a positive attitude to training tended to benefit more from it. However, they rightly noted the lack of evidence of causality; it might be that markers perceived the training's beneficial impact on their marking quality, and hence were favourably disposed to it. But conversely, it might be that those people who are by disposition more open to new ideas (and disposed to respond positively to questionnaires) would be more likely to benefit from a novel approach to training.

Finally, in this section, we summarise findings from Greatorex, Baird, and Bell (2002). Greatorex, Baird, and Bell's (2002) research participants indicated in questionnaire responses that: mark schemes, co-ordination meetings, discussion with other examiners, and scrutinising exemplar scripts were all perceived as useful in aiding markers to mark at the same standard (*ibid.*, at p. 5). The descriptive statistics that allowed this finding to be made were followed up by a comparison between individual markers' preferences for different elements of the standardisation process. In these pairwise comparisons, the researchers found the following differences:

- the mark scheme was judged to be significantly more useful than discussion with other examiners
- the mark scheme was judged to be significantly more useful than the exemplar scripts and associated marks for the candidates
- the co-ordination meeting was considered to be significantly more useful than discussion with other examiners
- the co-ordination meeting was considered to be significantly more useful than the exemplar scripts and associated candidates' marks. (*ibid.*, at p. 6)

Like the other researchers cited in the section, Greatorex, Baird, and Bell (2002) noted the contradiction between markers valuing elements such as the standardisation (co-ordination) meeting, but then that meeting not having demonstrable impact on marking quality in carefully designed research studies. They cautioned that this absence of evidence should not be construed as evidence of absence; that is, an argument for removing face-to-face meetings from standardisation processes (*ibid.*, at p. 12).

3.1.4 Cross-cutting issues in the standardisation literature

3.1.4.1 Community of practice

The notion of a community of practice has been discussed extensively in literature on marking reliability (see: Meadows and Billington, 2005, pp. 53 – 55). A strong community of practice (or shared understanding between professionals) has been assumed to be associated with reliable marking. However, the concluding thought from Baird, Greatorex and Bell's (2004) leading study of standardisation practices invites us to re-evaluate the importance of communities of practice:

> The result of the study with co-ordination meetings did not show that factors like ownership, discussion and a flat hierarchy affect inter-rater reliability. What remains to be identified is the relative importance of experience, ownership, feedback, discussion and the other factors

> that have been discussed in the process of examiners coming to a common understanding of a mark scheme and high levels of inter-rater reliability. Furthermore, the community of practice literature has great descriptive utility, but its prescriptive utility has yet to be established. How does one know whether a community of practice has already been formed and will fostering the features of a community of practice engender reliable marking? As yet, it is unclear whether particular features of communities of practice are necessary for reliable marking, or simply by-products of the community of practice. (Baird, Greatorex & Bell, 2004, p. 346)

Several of the studies that have followed Baird, Greatorex and Bell (2004), and which are reported at pp. 4ff, above, appear to point in the same direction; standardisation was carried out and there was little consistent evidence that face-to-face meetings provided higher reliability that online approaches. Further, there was no clear relationship between markers' perceptions of training initiatives and those initiatives' impacts (pp. 7ff). So, what are we to make of this evidence? Do we accept a position that communities of practice are less important than previously thought?

Perhaps expecting communities of practice to predict higher reliability is a category error. Perhaps that is not what they are for. Adie, Klenowski and Wyatt-Smith (2012) and Adie (2013) propose a vision in which technology is used to enhance shared teacher identities and to bind teachers into summative assessment systems. Tremain (2011, 2012) has pioneered research into the retention of examiners. The conclusions of this work are not yet clear, but it requires us to make sure that increasing use of technology do not lead to a shortage of markers.

3.1.4.2 Simulated intervention studies

The careful design of many of the studies included in this review is recognised. Nonetheless, as many of the authors themselves acknowledge, such care cannot transcend all limitations. In particular, some of the studies appear to be 'riding on the coat-tails' of pre-existing quality assessment practice. For example, Baird, Greatorex and Bell (2004) found that a mark scheme alone – without a standardisation meeting – was sufficient to maintain marker agreement. Similarly, Chamberlain and Taylor (2010) suspected the presence of a 'ceiling effect'. In other words, the participants in both studies were already skilled markers before being the subjects of the research. This casts doubt on the causes of effects reported in studies.

Extending the point, it is also important to note that only Knoch's (2011) research had a design that spanned several test administrations. There is a need for more longitudinal studies; to establish whether novel standardisation practices have a gradual impact – for example with markers drifting apart over time in terms of their internalised standards. Conversely, it is also possible that examiners who find novel approaches to standardisation and marking difficult at first, may over time become more familiar with them and thus reveal improvements in the reliability of their marking. So there is a real danger in concluding too much from 'one-off' simulated intervention studies, when long term changes in the quality of examination marking may be achieved in quite different ways if procedures for live examination marking are reformed and sustained over several diets of the same examination.

3.1.4.3 Statistical indices and traditions

The concepts behind, and the measures and applications of reliability theory have been extensively researched in recent years in the UK (He & Opposs, 2012). A few brief comments on the relevant analytical traditions used in research are apt.

Firstly, we recognise Bramley's (2007) careful categorisation of statistical indices to capture marking quality. He argues forcefully for conceptual simplicity in reporting indices of marker quality. We join Bramley in citing from the following passage:

> A study that reports only simple agreement rates can be very useful; a study that omits them but reports complex statistics may fail to inform. (Bramley, 2007, p. 27)

Further, we assert the importance of Bramley's insight that reliability is properly used to refer to properties of sets of scores, rather than individual scores, and that – in the latter case – the term 'agreement' is better than 'reliability'. Bramley expresses this as follows:

> The previous scenarios have concentrated on methods for assessing a single marker's performance in terms of agreement with the correct mark on an objective item (scenario 1), and agreement with the Principal Examiner's mark on a more subjective item (scenario 2). The term 'reliability' has been deliberately avoided. I would suggest we do not talk about the reliability of an individual marker, but reserve the term 'reliability' for talking about a set of marks. Thus reliability is a term which is perhaps best applied to an aggregate level of marks such as a set of component total scores. (Bramley, 2007, p. 26)

Related to this insight, we should also note the distinction made by several researchers cited in the standardisation section above between indices of 'reliability' and 'validity'. For example, Wolfe, Matthews and Vickers (2010) define an inter-rater reliability, and a validity coefficient and a validity agreement index, in the following manner:

> *Inter-rater reliability*: the correlation between the scores assigned by a particular rater and the average score assigned by all other raters in the project … . This index indicates whether a particular rater rank ordered examinee responses in a manner that is consistent with the typical rank ordering of those examinees across the remaining raters in the study, an index that is not sensitive to rater severity or leniency.
>
> *Validity coefficient*: the correlation between the scores assigned by a particular rater … and the consensus score assigned by scoring project leaders to those essays, another index that is not sensitive to rater severity or leniency.
>
> *Validity agreement index:* the percentage of exact agreement between the scores assigned by raters … and the consensus scores assigned by project leaders—an index that is influenced by several rater effects (e.g., severity/leniency, centrality/extremism, and accuracy/ inaccuracy). (Wolfe, Matthews & Vickers, 2010, p. 10)

It is worth noting that, given the hierarchical approach to standardisation contained in the UK Code of Practice (as outlined below at p. 54), it would be easy to (erroneously) conceive of Wolfe, Matthew and Vickers' validity coefficient and validity agreement index as reliability indices.

In addition to these strictures requiring clarity in thinking about exactly what indices of marking quality mean, we believe it is useful to compare the different measurement paradigms within which the researchers reporting findings on standardisation worked. We have adapted a categorisation system from that used by Baird et al (2013). We have noted the categorised approaches taken in studies selected for inclusion in the review as follows:

- Classical methods
- Multi-faceted Rasch measurement (MFRM)
- Generalisability theory (g-theory)

There are several comments that we can make on these categories. Firstly, we do not explain MFRM, g-theory (or indeed classical methods) in our report. This has been done exhaustively elsewhere (Baird et al (2013) would be a good starting point). Secondly, we acknowledge that there is a substantial degree of arbitrariness in such categorisations; advocates of g-theory and the Rasch model would emphasise their grounding in classical statistics, and some of the studies we have denoted as 'classical methods' below contain elements of the other approaches – for example Pell et

al's (2008) use of ANOVA has echoes of g-theory, whereas Wolfe and McVay's (2012) latent trait modelling is similar to Rasch modelling. Finally, we have not yet found any use of multi-level modelling (MLM) in the standardisation studies, although this method is used in several marking reliability articles (see: Baird et al (2013); Leckie & Baird (2011)).

The categorisations of analytical methods used are shown in Table 3, below, and a brief commentary follows the table.

Paradigm	Reference	Analytical and/or data collection technique
Classical methods	Bird and Yucel (2013)	Student's t-test Inter-rater reliability as SD around average marks awarded variation between markers and expert marker paper survey for opinions
	Chamberlain and Taylor (2010)	Marking accuracy (absolute mark differences) and consistency (rank order correlations)
	Greatorex and Bell (2008)	ANOVA analysis of differences between examiner marks and reference mark
	Pell, Homer and Roberts (2008)	General Linear models (a form of ANOVA) to compare: Student gender ● Assessor gender ● Assessor training ● The interactions between assessor training status assessor and student gender.
	Wolfe and McVay (2010)	Latent trait models to identify rater leniency, centrality, inaccuracy, and differential dimensionality; association between rater training procedures and manifestation of rater effects
	Wolfe, Matthews and Vickers (2010)	For rater quality: ● inter-rater reliability ● validity coefficient ● validity agreement index For rater perceptions: ● Two 15-item questionnaires ● Alpha and correlation indices
Classical methods/g-theory	Baird, Greatorex and Bell (2004)	● Analysis of actual and absolute differences ● variance components in g-theory for unbalanced data design ● relative error, phi and SEM under g-theory
G-theory	Johnson, Johnson, Miller and Boyle (2013)	G-theory compared consistency of markers just after standardisation and then at end of marking period.
	Xi and Mollaum (2011)	G-theory for overall reliability of ratings Questionnaire to gauge raters' opinions.
MFRM	Elder et al (2007)	Multifaceted Rasch analyses to compare levels of rater agreement and rater bias
	Fahim and Bijani (2011)	Pre- and post-interview data collection: Various outputs of the Facets software to study rater consistency and bias.
	Knoch (2011)	Facets output to study rater bias, and consistency questionnaire for rater opinions/reactions
	Knoch, Read and von Randow (2007)	multi-faceted Rasch measurement self-report questionnaire for opinions
	Wolfe and McVay (2012)	Rasch rating scale model, and indices for raters' severity, inaccuracy, and centrality

Table 3: Research methods used in studies selected for inclusion in standardisation section

The majority of studies employ what we have loosely referred to as 'classical method'. As the quote from Bramley (2007) above suggests the thoughtful application of 'vanilla' indices can provide important insights into the studied question. However, this is dependent upon the careful controlling of relevant sources of variation in experimental studies (see above, at p. 10). MFRM offers many tools for modelling matters such as: rater severity, consistency, central tendency, halo effect and so on. In contrast to some of the 'classically-based' studies, some MFRM investigations risk appearing very 'bitty', and providing a range of micro-level insights which are harder to accumulate to a broader understanding of the topic. The apparent under-representation of g-theory and MLM is somewhat surprising, and may be rectified in subsequent studies.

3.2 Mark schemes

3.2.1 Introduction

It is no coincidence that the mark scheme is seen as an essential adjunct of the item it is for. For example, the Code of Practice (Ofqual et al, 2011) requires that 'question papers/tasks and provisional mark schemes must be produced at the same time' (Ofqual et al, 2011, p. 19). Indeed, Pollitt et al (2008) and Bejar (2012) go one further and argue that the mark scheme, or at least an outline of it, should be produced before attempting to construct a task.

The main reason for this close association between the two elements of an examination task is that it is often difficult to distinguish between the two in terms of the effect on the accuracy of the marking. This is partly because the nature of the mark scheme is heavily dependent on the nature of the task it is describing, or in Pollitt et al's (2008) terms, the nature of the behaviours that the mark scheme describes will require tasks with very distinct features. For example, a decision to reward the possession of a very specific piece of knowledge calls for a task which will require the display of that piece of knowledge and as little extraneous material as possible; conversely, if the skills to be rewarded are the ability to build a balanced argument about, say, an historical event, then the set task will be very different.

The literature about the effect of mark schemes on marking accuracy therefore and quite rightly both goes along with and is often entangled with discussions on the effects of different questions on marking accuracy. Moreover, any discussion of these issues cannot be wholly separated from discussions of validity and much of the research reported here notes that any possible adjustments to schemes of assessment in the interests of reliability must have due regard to validity. Indeed, the key focus of Pollitt et al (2008) is on validity rather than reliability, but their process is about making the assessment of the construct concerned reliable so that other aspects of validity can be evaluated.

3.2.2 Terminology

Among other difficulties which occur when considering the effects of mark schemes on the accuracy of marking is that of terminology, and that applies to other aspects of question type and mark scheme type. Indeed, as Bramley, (2007) argues, it is important also to be careful when characterising marking accuracy, that one looks at the most defensible features to quantify it (cf. pp. 10ff, above). The most familiar categorisation of question types is into a binary distinction between closed and open or constrained and unconstrained. (See, for example, Sweiry, 2012.) However, Pollitt et al (2008) favour: 'constrained', 'semi-constrained' and 'unconstrained' to distinguish what are in effect differences in the nature of the responses, while Bramley (2008, 2009) uses objective, points-based and levels, drawing in the nature of the mark scheme to characterise the item type.

In addition some work has identified question features that go beyond the three level distinctions in Pollitt et al (2008) and Bramley (2008, 2009) and these are features that take in both aspects of the tasks and of their associated mark schemes. Bramley (2008, 2009) codes the items he investigates for features other than the item types described above. These include features of the question paper and response (answer space and the amount of writing expected) as well as coding features of the mark scheme (points-to-marks ratio, qualifications, restrictions and variants, and wrong answers specified. Pinot de Moira (2013) draws nine different distinctions in the features of levels-based mark schemes, each of which has the potential to affect marking accuracy.

The features are:

- number of levels;
- number of marks within a level;
- distribution of marks across levels;

- inclusion or not of quality of written communication;
- presentation in a grid-like format, to separate assessment by assessment objective;
- the inclusion of a mark of zero within the bottom level;
- the inclusion of indicative content within the levels;
- the order of presentation of the levels – low-to-high or vice versa;
- and the inclusion of advisory documentation on how to apply the scheme.

It is important to recognise that these nine features are not theoretical possible variants but variants that can be found within the mark schemes of a single awarding body. However, Pinot de Moira's (2013) work found that none of the features had a statistically significant bearing on marking accuracy, although there were some smaller, non-significant[10] effects, outlined in the table below.

Feature	Finding about reliability	Recommendation
Number of levels	None	Make commensurate with ability to describe clearly, limits of cognitive discrimination and weight within specification.
Marks within a level	None	
Distribution of marks	Non-significant improvement if marks evenly distributed	Distribute marks as evenly as possible across levels.[11]
Includes quality of written communication (QWC)	None	Assess QWC separately.
Separate assessment objectives	None	Design mark schemes with cognitive demand in mind. Clarity and conciseness are important.
0 in the bottom level	None	No recommendation
Indicative content within levels	Non-significant improvement if no indicative content	Design mark schemes with cognitive demand in mind. Clarity and conciseness are important.
Lowest or highest first	Non-significant improvement if lowest first	No recommendation
Instructions on how to use	None	Include clear and concise instructions for use.

Table 4: Features of mark schemes associated with reliability; findings from Pinot de Moira (2013)

There are similar variations in the way that different subjects and/or different awarding bodies present the information in points-based mark schemes. This usefully highlights the danger of referring to any category of mark scheme as if it perfectly expresses the nature of the mark scheme under discussion. Notably, while Pollitt et al (2008) call for much greater consistency, the best that they anticipate is a standardised approach to command words and the expectations they engender within a subject.

[10] We add this table to extract maximum information from the reviewed studies. Of course, since non-significant findings are often considered to amount to no findings at all. However, we retain the table as a source of useful guidance – so long as caveats around the significance of findings are recalled.
[11] This is strongly endorsed using theoretical examples in Pinot de Moira (2011a).

Only one of the features is the same as the most commonly investigated distinction between such mark schemes: analytical and holistic (see, for example, Çetin, 2011). Equally Suto and Nádas (2009) use Kelly's Repertory Grid techniques to identify question features that affect accuracy of marking using GCSE mathematics and physics questions, with several of the identified features relating to the nature of the response and the mark scheme.

How to characterise mark schemes has also led to different approaches. Perhaps the most important of these comes in Ahmed and Pollitt (2011), an article which offers a taxonomy of mark schemes, which can be applied to all forms of mark scheme from the purely objective to complex holistic levels-based ones. Moreover, the taxonomy is evaluative, suggesting that some forms of mark scheme are going to be better than others at achieving consistent marking. The nature of the taxonomy is explored further in Section 3.2.1.4 below.

3.2.1 Evidence found on main research themes

3.2.1.1 Interaction of mark scheme type and standardisation approaches

The advent of online marking has implications for mark schemes and standardisation. In particular, a frequently described feature of online marking is the potential to allocate different questions or sections of a question paper to different markers, often distinguishing the sections by the anticipated difficulty of marking, and then allocating only the hardest-to-mark to the most consistent of the marking team. There is evidence (cited in Meadows & Billington, 2005) that neither qualifications nor experience have any impact on the ability to mark very tightly controlled items accurately. Meadows and Billington (2010) show that even for shorter questions, involving considerable judgement, students with a PGCE were about as reliable as experienced examiners, while Royal-Dawson and Baird (2009) found that teaching experience was not necessary to mark even quite extended answers in English.

However, the tendency in these reports is to observe that the range of markers can be expanded to allow them to mark appropriate questions, subject to suitable training. However, there appears to have been no research on whether the form of the training should differ for different item and mark scheme types or indeed different markers. Baird, Greatorex and Bell (2004) suggest that there is little difference in marker performance, whether they have attended a standardisation meeting or not, nor whether the meeting was hierarchical or consensual in nature. Meadows and Billington (2010) show that there is an effect of training but that was not always in the expected direction in terms of accuracy, actually increasing the absolute mark differences for some sub-groups on specific questions.

Suto, Greatorex and Nádas (2009) found that marker training had a differential impact on marker accuracy for different questions in a question paper, greatly improving it in some cases, and even slightly worsening it in others. They do not, however, give details of the various questions, nor their associated mark schemes for which this is true.

What evidence there is in this area is therefore inconclusive. This is not surprising given the relative lack of investigation in the area and the fact that both mark schemes and standardisation meetings can vary considerably in nature.

3.2.1.2 Shared understanding of levels-based mark schemes facilitated approaches to standardisation

Levels-based mark schemes remain the instrument of choice for marking the vast majority of extended writing tasks in assessments worldwide. They are divisible into two categories, analytic and holistic and there has been some research into which of the two produces more consistent marking. Çetin (2011) randomly allocated novice raters to mark student essays either holistically or analytically and found that correlations were highest when both markers had marked it holistically

and lowest when one marker had marked holistically and one analytically. Pinot de Moira (2013) has the distinction between holistic and analytic as only one of the features that she identifies as having possible effects on consistency of marking. She argues, at least implicitly, for holistic schemes, since she claims that 'clear, concise and simple mark schemes are likely to elicit more reliable marking' (*ibid.*, at p. 9). Conversely, Ahmed and Pollitt (2011) and Pollitt et al (2008) would appear to favour analytical mark schemes, again by implication, since such schemes are clear as to the principle of how to apportion credit between different skills/assessment objectives. Here too, however, there is little evidence in the research as to which approach is better in terms of creating a consistent and shared understanding of the scheme.

The International Baccalaureate Organisation (IBO) is currently proposing to change the mark scheme for the essay element in the Theory of Knowledge within the IB Diploma from analytical to holistic (IBO, undated). Before they commit to this they have carried out a trial of the two mark schemes. The trial was relatively small scale (involving 16 markers and 40 essays). The grades the markers gave to the essays were compared to definitive grades (defined as the grade agreed by the senior examining team, under each scheme). The findings show that these *definitive* grades under the new (holistic) criteria were much more bunched than under the old criteria with 80 per cent of the essays getting a C or D under the new scheme as opposed to 55 per cent under the old. However, the distribution of grades awarded by the *markers* was more similar across the two schemes (with over 70 per cent being awarded a C or a D under both). The markers also awarded two-and-a-half times as many A grades under the old scheme and about half as many E grades, a significant shift toward the lower grades. The grade arising from the marking under the new criteria was more often the same as the definitive grades than was the case with the old criteria (over 50 per cent as opposed to about one third) and there were no cases of extreme differences in outcomes (± three grades) where there were four per cent using the old criteria. It should be noted that the latter results, in part, arise as a result of the much greater bunching in the definitive grade within the new scheme. Interestingly, the report comments: 'it is clear that further work is needed if a decision is made to use the new criteria during a live examination session, especially if it is during the current curriculum cycle' (*ibid.*, at p. 5).

It is clear that the IBO has reasonably high expectations of levels of agreement between markers and the definitive mark, although this is not quantified. But it is not clear from any research what aspects of levels-based mark schemes best facilitate accuracy nor how they best feed into standardisation.

3.2.1.3 Features of questions and mark schemes that can affect accuracy

Bramley (2008, 2009) investigated levels of agreement (defined as the level of exact agreement between the examiner mark and the team leader mark during the sampling process) for a range of question types and mark scheme approaches across a range of 38 public examinations. He found that each of the features had an impact on the accuracy of marking. In most cases this was in the predictable direction (e.g. there was consistently less agreement the greater the maximum mark for an item, although there is an anomaly for objective items with a maximum mark of three). He also found that, for objective items, the presence of either qualifications, restrictions and variants[12] or of possible wrong answers in the mark scheme reduced the levels of agreement and that levels of agreement were higher for points-based than levels-based mark schemes for relatively low tariff items. The situation was reversed for high tariff ones. The main findings are summarised in the table below:

[12] Defined as 'the presence of additional marking rules or guidance' in Black, Suto and Bramley (2011). Bramley (2009) makes clear that this includes items which include error carried forward.

Feature	Finding about reliability
Maximum mark	The level of agreement declines as the maximum mark increases.
Mark scheme type (objective; points-based; levels-based)	For a given maximum mark, objective items are more accurately marked than points-based.
Answer space	There was slightly higher agreement the smaller the available answer space, for a given maximum mark.
Writing	The writing category was affected by the type of answer expected.
Points-to-marks ratio	There was greater agreement when the number of valid points was the same as the number of marks than for when there were more.
Qualifications, restrictions and variants (QRVs)	There was generally greater agreement for objective items when there were no QRVs. The situation was often but not always reversed for points-based mark schemes.
Wrong answers specified	There was greater agreement for objective items when no wrong answer was specified. The situation was less clear in points-based items.
Points vs. levels	For relatively low tariff questions, points-based mark schemes were associated with greater marking accuracy. The situation was reversed for higher tariff items (max mark greater than 10).

Table 5: Impact of mark scheme features on reliability found by Bramley (2008, 2009)

Black, Suto and Bramley (2011) extend Bramley's work (2009) to include possible features of candidate response which one might expect not to be relevant (for example, legibility of handwriting, or answers that go out of the prescribed area). They also observed examiner meetings allowing some further potential features of the marking process to be included in the analysis (time taken by the senior markers to agree a mark for each item; level of contention about the agreed mark; level of democracy in determining the agreed mark). They also made use of the online marking system to investigate marker differences, comparing marks given by assistant markers to seeding items. This is a key difference from Bramley (2009) since these items are being second marked blind whereas the marks Bramley (2009) used, those arising from the sampling process, involve the team leader seeing the marks originally awarded by the assistant marker.

In terms of question and mark scheme features, Black, Suto and Bramley (2011) tend to confirm the findings in Bramley (2009), although points-based mark schemes consistently produced higher levels of marker agreement than levels-based ones.[13] What is particularly interesting in this study is the examinee response features which were most strongly associated with marking accuracy. The first – whether the response was typical or atypical – is perhaps not surprising and takes one back to the arguments of Ahmed and Pollitt (2011) and Pollitt et al (2008) about the need for a mark scheme to be clear about the principle behind the award of credit rather than simply trying to anticipate the outcome space. The other two – presence of crossings out etc. and writing outside the designated area for the answer – are clearly not construct relevant, but both reduced the accuracy of the marking. They also found that discussion time and level of contention were both significant predictors of marking accuracy (in the expected direction) perhaps suggesting that the greater the

[13] They do not, however, make clear what the maximum marks for the various questions were. The crossover point in Bramley (2009) was about 10 marks, and it is easy to see how checking so many different possible points in an answer against the mark scheme could lead to error.

cognitive demand required to understand the demands of the question and thus apply the mark scheme, the harder it is to mark accurately. The authors also present a set of possible actions designed to improve marking accuracy, with an indication of the likely impact on marking accuracy and any other possible effects, such as an impact on validity.

The finding of construct-irrelevant factors impacting on marking accuracy is in line with the findings of Smart and Shiu (2010) and Rezaei and Lovorn (2010) who deliberately manipulated two essays so that one was well written in technical terms but failed to address the task while the other was fully relevant but contained a number of technical errors of spelling and grammar. Participants scored each essay twice, once without a mark scheme and once with (essentially the holistic versus analytical divide). The mark scheme was designed to focus assessment on relevance rather than technical accuracy. Participants were themselves divided into two groups, one of whom (education students) were assumed to be familiar with the application of mark schemes; the other (business and marketing students) who were assumed to be unfamiliar. None of the participants was trained in using the mark scheme in question, but it is interesting that the marking of the group which was assumed would be more familiar with marking to a rubric was the less accurate of the two for both approaches to marking.

Findings were consistent across both groups: the use of the mark scheme significantly deflated the marks awarded (by 10 percentage points or so) and the education students rated the superficially more correct essay higher than the more relevant one, both with and without the mark scheme. The business and marketing students rated the more relevant essay more highly than the other one without the mark scheme but the two about the same when using the mark scheme. In other words, the introduction of the rubric did nothing to help the markers focus on the appropriate features of the work. The question of construct irrelevant factors will be returned to in section 3.2.1.4, below.

Suto, Greatorex and Nadas (2009) found that the probable complexity of the marking strategy had a strong relationship with marking accuracy, with those questions requiring an apparently complex strategy being less accurately marked. This was true whether or not participants had been trained, although their accuracy did improve as a result of training, especially in physics.

The issue of the complexity of the marking strategy required underpins most of the findings in this section and will be further explored below.

3.2.1.4 Different question/mark scheme types and marking strategies

Part of the key to an effective mark scheme must lie in anticipating and understanding the thought processes a marker may use when assessing a script and then seeking to create a scheme that will, as far as possible, direct those processes towards a consistent and valid evaluation of an answer.

The work of Crisp (2007, 2008a, 2010), Suto and Greatorex (2008) and Suto, Greatorex and Nadas (2009) using verbal protocols has enabled considerable exploration of the kinds of strategies markers use to come to decisions about marks and how these relate to the mark schemes. This has led to several ways to characterise the marking process and how it relates to the mark schemes: the identification of apparently simple and apparently more complex strategies; suggested taxonomies for the different types of strategies markers employ; suggested ideas for the stages (sometimes iterative) the marker goes through in arriving at a decision; and evidence of several questionable factors that seem to be active in the marking process.

Suto and Greatorex (2008) used examinations in mathematics and business studies to investigate the processes markers went through when marking questions using points-based (mathematics) and levels-based (business studies) mark schemes. After a re-familiarisation process involving standardisation and some further marking; the markers, all of whom were experienced in marking the relevant syllabus, were invited to talk through their thought processes as they marked a set of five scripts. Transcripts of the resulting tapes were coded and analysed, and they resulted in five

strategies being identified: matching, scanning, evaluating, scrutinising and no response. Essentially, markers seemed to be checking what they found in an answer against the mental model they had formed of the mark scheme, a process also described in Bejar (2012). Examples of each strategy could be found for both subjects, participants found the labels convincing in interviews and other examiners in the same subjects felt that the information could be very useful in the item and mark scheme writing as well as the standardisation of marking.

Crisp (2007) used an AS level and an A level geography paper. It is not explicitly stated in the report, but it appears that most, if not all, of the questions in the papers were marked using a levels-based mark scheme. Crisp does however note that 'variations between marking short-answer questions and marking essays were apparent in certain phases of the model' (*ibid.* at p. 17). As with Suto and Greatorex (2008) participants, all of whom were experienced markers for one or both of the papers, first had a re-familiarisation phase in which they marked a selection of scripts, the first batch of which was reviewed by the Principal Examiner for the paper. Marking was broadly in line with that done operationally, but with a slight tendency to severity.[14] Markers each then attended a meeting where they marked some scripts in silence, then a further set of scripts while thinking aloud and then were interviewed.

Behaviours were categorised into reading and understanding; evaluating; language; personal response; social perception; task realisation; and mark. There were some variations in the frequency with which different examiners exhibited these behaviours and they also varied in frequency according to the nature of the question (short answer, medium length answer or essay). There was also some possibility that the pattern of behaviours could give some clue to the accuracy or otherwise of a marker but numbers are far too low for this to be other than speculation (and the possibility that thinking aloud interferes with the marking process differentially for different markers).

Crisp also speculates about the behaviours and how they might suggest a possible model of the marking process. This is something she takes further later (Crisp, 2010) when she re-analyses the data with a view to becoming clearer about the marking process. The model she comes up with is a five-stage process, with the first and last stages not always occurring and the three intermediate stages being essentially iterative. She suggests there may be a prologue stage of thoughts arising before actually getting into the answer, perhaps orienting themselves to the question followed by phase 1: reading and understanding with evaluation. This is a mix of specific reading strategies: scanning text for specific items, paraphrasing etc.: coupled with concurrent evaluation – identifying good points, relating the answer to the expectations of the mark scheme, etc. This phase matches the model of the answer being created with other models, whether those suggested by the mark scheme, those from other candidates or simply the marker's own expectations. Evaluations also included qualitative references to language, presentation and handwriting and various affective and social dimensions. The next phase is overall evaluation of response, where the marker attempts to summarise thoughts about the answer (again often using various forms of comparison) before moving on to phase 3: the award of a mark. The marker often loops between these two phases, particularly in the case of extended answers. There is also sometimes an epilogue, where the marker continues to comment on an answer even after the award of a mark. Clearly understanding the process has potential benefits for design of assessments, construction of a mark scheme and marker training.

Elliott (2013) also used verbal protocol analysis in her investigation of marker behaviour. She found that while there was frequent reference to the mark scheme, especially during standardisation and the early phases of marking, there was rather more evidence of comparisons to various models: the

[14] This is a familiar pattern when marking is done away from operational conditions.

sample or anchor scripts; other scripts that the marker has already marked; and imagined or idealised scripts for the various levels of response. She also found much more evidence of the epilogue phase than Crisp (2010). Indeed, during such a stage, markers would sometimes return to previously marked scripts and amend the mark awarded in the light of the mark the current script is judged to deserve. It seemed that the main objective was to place the scripts in the right rank order rather than direct reference to the mark scheme.

Suto and Nádas (2009) took a different approach to investigating the demands of the marking process, using Kelly's Repertory Grid to try to identify relevant question and mark scheme features. They focused on mathematics and physics examinations. They related the constructs identified by principal examiners in the two subjects to information about marking accuracy, question difficulty and apparent cognitive marking strategy. In both subjects only some of the identified constructs did actually relate to marking accuracy but in both cases predictions by the Principal Examiners to how easy a question was to mark did generally (although not perfectly) relate to marking accuracy. One feature in mathematics which was associated with reduced marking accuracy was error carried forward. In both subjects the use of context affected marking accuracy as did alternative answers in mathematics/flexible mark schemes in physics. In physics there was a cluster of features around reading and writing which also affected marking accuracy. Again, knowledge of these features is clearly of potential value in question and mark scheme design, but as the authors point out 'removing questions with features associated with low accuracy could create a different problem: the reduction or elimination of important and valid constructs from an examination' (*ibid.*, at p. 356).

3.2.1.5 Subject-specific effects

One of the issues that consistently arises with investigations into examinations is the level of generalisability of any findings. Can an outcome about one subject – or indeed at one level – be extended to other subjects and/or levels? For example, Suto and Nádas (2009) concentrate on GCSE mathematics and physics in their work; Crisp (2007, 2008a, 2008b, 2010) uses data arising from one AS and one A2 paper in geography. Even though Black, Suto and Bramley (2011) and Bramley (2009) look at a much wider spectrum of subjects and levels, there is a warning that one finding 'should be treated with some caution however, because the high-mark levels-based items were strongly clustered in particular units (subjects)' (Bramley, 2009, p. 20)[15]. Similarly, while Pollitt et al (2008) call for much greater consistency in the use of command words, and how they translate into expectations in the mark scheme, they remain clear that 'at best, it might be possible to define a command word as it is normally used in GCSE in one subject' (*ibid.,* at p. 77).

It is a fact that each subject (and, as Pollitt et al (2008) recognise, sometimes each specification within an awarding organisation) has its own preferred approach to assessment. This includes the pattern of question types used possibly varying by level. (Crisp (2007) chose an AS geography paper because it has a mix of short- and middle-length items and the A2 one because it comprised essays. This is not necessarily the pattern with all A level geography examinations.)

This not only constrains the generalisability of any findings but also constrains the likely scope of any research. Thus, when Coniam (2009) wished to investigate the possible effect of mode of marking on the assessment of extended writing, the focus was on the longer of two questions in the English language examination. Similarly, O'Donovan (2005) went to a unit from an A-level Politics syllabus to investigate the marking of essay-based responses. Conversely, in the mathematics and physics papers that Suto and Nádas (2009) used, overall questions were relatively low tariff and were often further divided into parts carrying one or two marks. In the end it is therefore difficult to disentangle subject specific effects from question/response effects. Even where one can disentangle these effects there is of course no certainty that a marking approach that works well with one group

[15] The subjects are unnamed.

of examiners will generalise to multiple groups of examiners, with their own characteristics, backgrounds and identities and priorities.

3.2.1.6 Alternative approaches to a mark scheme

The literature about marking has one additional strand, which potentially removes the necessity for a mark scheme altogether. Pollitt (2012) argues for a method of Adaptive Comparative Judgement as a way of assessing portfolios or other significant bodies of work holistically. The technique builds on two existing models, the first being that of Thurstone Pairs as a way of capturing holistic comparative judgements. The second is the idea of the computerised adaptive test which uses the information about the candidate taking the test from performance on items already taken to select subsequent items as being most helpful in increasing the precision of the measurement of that candidate's ability. The system uses Rasch analysis[16] to build up an early picture of the relative success of each piece of work and matches it to other pieces of very similar standard to create an overall rank order. Pollitt reports very high levels of reliability with the method. The analysis is also capable of detecting and thus allowing for examiner bias.

As Pollitt (2012) points out, comparative judgement based on Thurstone's principles is already the preferred method of investigation in comparability studies and was also used to investigate the accuracy of standard setting judgments of candidate's work by Gill and Bramley (2008). They used scripts from A level examinations in physics and history. Importantly they found that comparative judgements between scripts were much more accurate than absolute judgements (in terms of the grade a script deserved). It should be noted that the right mark and grade in these cases was always the one the script had received as a result of the application of the mark scheme. Overall, therefore, the rank ordering that would arise from a process of comparative judgement would not be greatly different from that produced by the marking process.

Jones, Swan and Pollitt (in press) used comparative judgement to assess mathematical problem solving, which does not lend itself easily to traditional methods of assessment in mathematics. They also investigated the method when assessing existing GCSE mathematics examinations which are highly atomistic and thus not particularly suited to holistic forms of judgement. They reported encouraging levels of inter-rater reliability and of consistency of outcomes with those in the original examination.

The very fact that so much of the marking process necessarily involves making comparisons, whether the marker is directly comparing a mental model of an answer with a mental model of the requirements of the mark scheme, or with other answers of known merit, means that the comparative judgement method is very attractive, the more so since it seems to produce reasonably accurate outcomes. However, to move to such a process operationally would require an enormous act of faith, given the risks attendant on it.[17]

3.2.2 Cross-cutting issues in the mark schemes literature

3.2.2.1 Theoretical approaches

One of the noticeable features of much of the recent work looking at marking and how to make it more reliable is a strong theoretical dimension. Within this work there are several clear strands. The first is an attempt to root the marking process and the items and mark schemes which should drive it within current theories of psychology and particularly the psychology of decision making. Because they are predominately theoretical, these papers often offer little by way of evidence to back up their arguments, but they often provide relatively straightforward means by which their

[16] As a mathematical enhancement of Thurstone's models.
[17] It is also hard to see how the enquiry about results process could easily be incorporated into such an assessment system.

claims could usefully be investigated, and if they are substantiated offer the potential greatly to improve the process of marking either through the assessment instrument itself (in which, as already argued, the mark scheme is at least as important as the item itself) or through marker training.

One of the principal methods used to develop the theoretical work has been the use of verbal protocol analysis (Crisp, 2007, 2008a, 2008b; Suto & Greatorex, 2008) to gain some insight into the actual process of awarding marks and to relate that to theories of decision making. Think-aloud protocols have also been used on a more pragmatic basis to investigate how particular examiners whose marking had been found questionable approached their task (Smart and Shiu, 2010).

Getting markers to voice their thoughts as they mark candidate work has allowed two separate developments which have real potential for improving the way markers apply the mark scheme, provided the ideas are thoroughly validated, which will be explored further below. The first of these has been to identify the five different strategies for determining a mark (Suto & Greatorex, 2008): matching, scanning, evaluating, scrutinising and no response, with the first, the fifth and to a large extent the second associated with what they identify (using Kahneman & Frederick, 2002) as system 1 thought processes which are quick and associative, and the third and fourth requiring more system 2 processes, which are slow and rule-governed. It is easy to see how knowledge of these strategies would prove useful in thinking about how best to control markers' thinking. In particular, it is vital to ensure that the mental models against which the markers are matching, scanning, etc. are firmly rooted in the mark scheme.

The second development has come from Crisp (2007, 2010) where she advances a model in which the marking process is divided into five stages, although not all will necessarily be used in arriving at a particular marking decision. In fact, Crisp (2010) reports that some of the stages are virtually indistinguishable during the marking of relatively short answer items. The stages are prologue, where the marker may have some thoughts before reading the answer; reading and understanding; overall evaluation; mark decision; epilogue, where there may be thoughts arising after the marking decision. The phases may also be iterative, especially phases 1 and 2. Again it is easy to see the potential of this model for helping understand and improve marking, especially as it helps to identify the stages where construct-irrelevant features are most likely to come into play. Consequently, it should be possible to use this understanding to minimise the effects of such construct-irrelevant features both in the mark scheme and standardisation process. CECR (2012) give an outline of how frame of reference training (essentially standardisation in using a levels-based scoring rubric) needs to try to predict and pre-empt construct irrelevance.

Moreover, by interviewing participants in the studies that have given rise to these analyses, the authors have, to some degree, validated their claims. However, there are also inevitably some question marks over the findings, and their generalisability. First, there is the issue of whether being asked to think aloud while marking actually alters the marking process itself. The papers that make use of verbal protocol analysis claim not (see also Crisp 2008b) but the case remains unproven, especially since in Suto and Greatorex (2008) three of the five the mathematics markers were unable to complete the process in the time allowed, although it was relatively generous. For a start, it seems quite probable, that the mere act of verbalising a system 1 thought has the potential for making it system 2.

In addition, the studies are highly resource intensive and thus necessarily small scale. Suto and Greatorex (2008) looked at GCSE mathematics and business studies[18] (arguing that they covered both a points-based approach and a levels-based one) with Crisp (2007, 2010) looking at AS and A2 papers in geography (which offered a mix of relatively short and extended answer questions).

[18] In fact GCSE mathematics questions have mark schemes that are close to objective, and marking of many GCSE business studies questions is points-based.

Encouragingly, many of the voiced comments showed that a key reference point for the markers was the mark scheme. Interestingly, in the work by Smart and Shiu (2010), which was investigating the processes used by markers who had been found relatively unreliable, a major concern was the extent to which they drew on judgements which were outside the mark scheme, although sometimes related to the key categories within it. However, it is not clear whether the same would be true of markers who had been found to be reliable (that is the next phase of the work). More importantly, it is clear from the English work using voiced protocols that a number of construct-irrelevant factors are at play. In particular, Crisp (2008a, 2010) found many examples of affective aspects, with markers reflecting on personal characteristics of the candidates and on the teaching that they had received. And Rezaei and Lovorn (2010) produce evidence that markers pay little regard to the contents of a rubric in arriving at a rating decision, with no real difference between those with experience in the use of a rubric and those without.

The second major strand in the theoretical work has been in using the theory of outcome space to develop a taxonomy of mark schemes. This forms the basis of the work by Pollitt et al (2008) in which they argue that the purpose of an effective mark scheme is not only to anticipate as many of the possible responses to a task and to give a clear indication of how to reward them but also to make clear the underlying principle behind that advice to allow markers to judge what credit to give for unanticipated answers. They illustrate their arguments with an extensive survey of GCSE question papers in business studies, design and technology, and geography, supplemented by evidence from examiners' reports. Their taxonomy identified four levels of mark scheme:

- Level 0: a mark scheme that gives no real help to markers, such as 1 mark for suitable answer
- Level 1: a mark scheme which is essentially just a model answer, with no guidance as to what to do with answers which don't match the model. Note, this applies even to relatively constrained questions where the mark scheme doesn't make clear exactly how much of an answer is required for a mark.
- Level 2: a mark scheme which attempts to provide information about the entire outcomes space. This can be divided into two types
    - Type 2a: lists points acceptable for a mark. Unless the list really is exhaustive (including for example all acceptable spelling variants) there will still be cases where examiners differ in their judgement.
    - Type 2b: a mark scheme which lists both acceptable and unacceptable answers. The same difficulty arises as with type 2a3: the list is unlikely to be exhaustive.
- Level 3: a mark scheme which, in addition to any indication of acceptable/unacceptable answers, makes clear the principle behind such decisions.

Pollitt et al (2008) do not suggest that every type of item should have a top-level mark scheme, but they are able to show how problems may arise even with relatively constrained items, unless the mark schemes make clear exactly what is expected before credit can be awarded. And, of course, the more complex the response the question is seeking, the more demands it places on even a level three-mark scheme to make clear the key principle without over-complicating it.

Jeffery (2009) also produced a taxonomy of mark schemes, albeit a much more focused one, as a result of her investigation of the direct assessment of writing in US assessments. In analysis both state-wide and national assessments, she identified six different categories of writing prompt: persuasive, argumentative, narrative, explanatory, informative and analytic. She also identified five different types of scoring rubric (essentially levels-based mark schemes): rhetorical, formal, genre mastery, expressive and cognitive. However, there was little real correlation between prompt and rubric type, with rhetorical ones associated with all six types and genre mastery with every type

except informative. Only the cognitive rubrics were associated with only one type of prompt (argumentative). She suggests that the difference between state-wide and national estimates of writing proficiency, with states consistently producing higher outcomes, may lie in the different expectations implied by different rubrics, making it hard to prepare students for both.

The third theoretical strand is largely evidenced by Pinot de Moira, and it involves mathematical modelling to investigate possible effects. Pinot de Moira (2013) explores levels-based marking schemes in this way, while Pinot de Moira (2012) explores levels-based mark schemes from a slightly different angle, showing how the way marks are distributed across the levels has the potential to skew the marks. Pinot de Moira (2011) models the effectiveness of a particular structure of question paper and mark scheme showing that it does not function effectively as a discriminator. She argues that each mark should be of equivalent value in terms of the step on the trait they are measuring, and therefore equally easy to achieve. Then, using a hypothetical example, she shows that the way even low-tariff questions where this isn't the case may affect the candidate rank order. She goes on to use an example of a unit where for some of the items, several of the marks are much less easy to obtain than would be expected from the overall mark distribution. In particular, one type of mark scheme, which might be represented as a cross between levels- and points-based, significantly underused several of the marks in the mark range and as she argues 'any item with under-utilised marks has the potential to limit discrimination between candidates' (*ibid.*, at p. 12).

3.2.2.2 The impact of online marking on research into mark schemes

One of the other interesting features of the recent studies is that the nature of empirical research on UK examinations has to an extent changed. In the past, attempts to investigate examinations empirically have almost inevitably been through simulated intervention studies. Considerable effort is usually made to make conditions as similar to the operational environment as possible, but this is inevitably limited in its success. Scripts are cleaned photocopies; examiner training does not have the same structure nor personnel; and perhaps most significantly markers are not operating in the knowledge that their decisions have the potential to affect a candidate's future. Conversely, simulated examination marking studies do allow for the opportunity to vary circumstances and, to some degree, control for variables so that attention can be directed to the issues under investigation.

The advent of online marking means that quite a lot of investigations, especially quantitative ones, can be carried out under operational examination conditions. This will often be large scale (it can cover all examiners and candidates) and the data are the operational data. Moreover, should any scripts require scrutinising as part of the process, these are essentially the same as those used operationally: scanned images, possibly in hard copy. The existence of large-scale data sets in and of itself allows for some control of variables (for example to study gender effects or markers with differing levels of accuracy) but cannot be allowed to interfere with operational success. It would be extremely unlikely to have work marked according to two marking schemes (for example a holistic one and an analytic one, as with Çetin (2011)) unless every script were marked according to both, with some agreed process for awarding a final mark. And that would create immense pressure on resources.

There remains therefore a clear place for the relatively small-scale simulated intervention studies, but it is to be hoped that the use of item-level data and evidence from blind double marking of seeded items be extended to look carefully at the effect of different types of marking scheme in different subjects. In particular, wherever an anomaly arises in the accuracy of marking a specific item, the item and its mark scheme should be scrutinised to see if the cause can be determined.

3.3 Online training and virtual meetings

As the final arm of the findings, we present an impressionistic review of webinars, technology-mediated meetings/communication and related issues. We do this as background for the prior two

sections, which are more particular to marking high-stakes examinations, and which, as a rule, report the empirical outcomes of research studies more specifically and precisely. This third set of findings seeks to complement the other two parts; by bringing some insights from across technology research; to consider how online meetings are done in general, and to try to show how such general challenges might apply to the specifics of standardisation and mark scheme use in a rapidly-changing technology environment.

Many organisations now have geographically dispersed teams working on joint projects, and so rely heavily on digital technology for team communication. The reported research covers a wide range of activities, from 'traditional' online learning to group dynamics to program evaluations. Few of these topics have any more than a tangential relevance to online standardisation. However the research around the use of technology in task focused discussions, group decision making and argumentation do have appreciable overlap.

Awarding organisations use a range of technologies for online standardisation. Much of the research in computer-mediated communication/conferencing/collaboration refers to text-based communication: email, discussion forums, chat, etc. (e.g. Luppicini, 2007; McAteer & Harris, 2002; Veerman, Andriessen & Kanselaar, 1999; Warkentin, Sayeed & Hightower, 1997). Whilst the expectation is that awarding organisations will rely predominantly on virtual meeting technology (i.e. web conferencing/collaboration software with shared desktops) to run online standardisations, some may use email or other text-based tools for pre-meeting discussions, or may include tools such as chat messaging as part of their virtual meeting system. Some aspects of the CMC research may therefore be relevant.

The literature commonly states perceived advantages and disadvantages of CMC when compared to face-to-face meetings (e.g. McAteer & Harris, 2002). Advantages quoted are often related to convenience and cost savings, but also include a perceived democratisation of the process (all participants on an equal footing, less likely to be biased by a dominant personality) (Veerman, Andriessen & Kanselaar, 1999), and for text-based systems the fact that the discussions are effectively archived and can be reviewed at any time. Disadvantages include the loss of paralinguistic information (body language, facial expressions, other non-verbal clues), and the feeling among participants that the medium is socially opaque (i.e. they do not know who or sometimes even how many people they may be addressing), and that misunderstandings may be harder to overcome (Veerman, Andriessen & Kanselaar, 1999; Thompson & Coovert, 2003). A review of computer-mediated communication research for education reported mixed research findings (Hinds & Weisand, 2003). Although there was some evidence that CMC groups interacted less than face-to-face groups, they tended to outperform face-to-face groups in task focused interactions (*ibid.*). Lantz (2000) also reports that collaborative virtual environment (CVE) meetings are more task oriented than face-to-face meetings. Earlier research on text-based systems had found that it takes longer to reach a decision using digital communications than in a face-to-face meeting (*ibid.*; Warkentin, Sayeed &Hightower, 1997).

More recent work investigates more sophisticated technologies such as web conferencing, web collaboration[19], and virtual reality (e.g. Erickson et al, 2010; Chen et al, 2007; Suduc, Bizoi & Filip, 2009). While such research is more sparse than for CMC (Erickson et al, 2010; Vailaitis et al, 2007), it is probably more relevant to online standardisation going forward.

[19] Web conference software typically provides features for document and file sharing, shared desktop access, and other electronic forms of communication that allow data to be shared, edited and copied during the web meeting (Suduc, Bizoi & Filip, 2009). Web collaboration software is similar but supports a many-to-many model with a range of two-way communication options. However as software vendors continually add features, the boundaries between web conferencing and web collaboration are blurring.

There is some support in the literature for the view that web conferencing or collaboration systems, comprising a shared desktop facility and possibly video conferencing, are more likely to be effective in virtual meetings (Hinds & Weisand, 2003). However while Erickson et al (2010) found that web conferencing with shared desktops rated more highly with users for information sharing than text-based communication, face-to-face meetings scored more highly still. Web conferencing with shared desktops are perceived to outperform virtual 3D systems (such as Second Life) in terms of information sharing and results-oriented work, but virtual 3D systems scored higher for social engagement. In all categories face-to-face meetings scored highest.

However, comparisons between different technologies have limited value in isolation. The key question is whether a particular technology can improve the performance of a task. Goodhue (1995) refers to this in terms of a 'task-technology fit' (TTF) model. The general concept is that, for a technology to be useful in enhancing performance, it must fit the nature of the tasks that it is designed to support. A good fit between task characteristics and technology characteristics will result in higher utilization and better performance (Turban, Liang & Wu, 2011). The concept of 'task technology fit' is, in the context of online standardisation, an important one; it is tempting to draw generalised conclusions from a meta-analysis of multiple studies on online training and virtual meetings (e.g. McAteer & Harris, 2002; Stahl, Koschmann & Suthers, 2006; Lantz, 2000), but to do so assumes that the mechanics of the virtual meetings are homogenous and the technology is commoditised. Neither of these are appropriate assumptions in the context of the current research; online standardisation has particular characteristics which distinguish it from, say, management or design meetings. Similarly, digital communication systems vary greatly; technology can be a barrier or an enabler, depending on the process requirements and the available technology features. This is recognised in the literature by researchers who report that the success depends very much upon implementation (Luppicini, 2007; McAteer & Harris, 2002; Hinds & Weisand, 2003). Best results from virtual meetings can only be achieved if the technology is able to directly support the required processes of the task (Chidambaram & Jones, 1993; Hinds & Weisand, 2003; Chamberlain & Taylor, 2010), and adequate technical support and training need to be provided for successful on-going implementation (Vailaitis et al 2007). As technology products develop, vendors tend to add features and so the products from leading suppliers tend to converge[20]. Even if the software has the features required to support online standardisation, there is no guarantee that the system is being used to best effect. As Suduc, Bizoi and Filip (2009) state:

> When web conferencing systems don't reach their full potential, it's likely because the participants aren't sharing information, not because the technology has failed. (*ibid.*, at p. 11)

There are a number of other issues which emerge from our review of the wider question of the effectiveness of online training and virtual meetings.

The first is that even where a positive result is being reported, several studies reported that not all users were equally comfortable with the technology (McAteer & Harris, 2002; Hamilton, Reddel & Spratt, 2001; Vailaitis et al, 2007; Erickson et al, 2010). The concept of teachers' (alleged) 'technophobia' is well discussed in the literature, e.g. Wood et al (2005); Lloyd & Albion (2005), and this may be a concerning factor for online standardisation. It may not be realistic to assume that all examiners are sufficiently comfortable with the technology to contribute fully to an online standardisation process, and this may influence their approach (Vailaitis et al, 2007).

The second issue is around user perceptions and satisfaction. Even where a study reports a positive result, there are often reservations reported by users. Loss of interaction compared to face-to-face meetings (Hamilton, Reddel & Spratt, 2001), lower perceptions of group cohesion and effectiveness

[20] Suduc, Bizoi and Filp (2009) report a very 'narrow margin between vendors' for web conference software.

(Luppicini, 2007), and concern about a potential loss of a community of practice (Chamberlain & Taylor, 2010) are all reported. However these findings are often mixed. Luppicini (2007) also reports that online participants reported lower levels of evaluation apprehension and peer influence than face-to-face participants. Perhaps the conclusion here is that any survey of user satisfaction needs to be considered in the context of the process as a whole.

The third issue is around the longitudinal use of virtual meetings. The importance of shared understanding in standardisation is recognised (Chamberlain & Taylor, 2010; Brindley, 1998), as is the importance of communities of practice (Elwood & Klenowski, 2002; Baird, Greatorex & Bell, 2004). Indeed, a section of this report explores this issue and its implications for standardisation practice (above, at p. 9ff). But in the paragraphs that follow, we look at the same issue through the slightly wider lens of general research on technology-facilitated meetings.

Shared understanding can have a significant impact on the ability of teams to perform well; with a shared understanding of the processes to follow and the meaning of information and data presented, team members are more likely to work effectively as a group and also more likely to be satisfied with and motivated by the process (Hinds & Weisand, 2003). There is also evidence from the literature that trust among team members affects the performance of virtual teams (Paul & McDaniel, 2004). However shared understanding and trust are more difficult to generate in virtual teams where team members rely heavily on technology to mediate their interactions and have less opportunity to talk through problems, share perspectives, and get feedback (*ibid.*). Without such communication, misunderstandings are more frequent and more difficult to resolve (Donnellon, Gray & Bougon, 1986). Hinds and Weisand (2003) list having the opportunity to learn about each other over time, communicating and sharing information, and developing a team spirit as factors which contribute to shared understanding, and stresses the importance of having occasional face-to-face meetings between virtual teams in order to establish rapport and common ground. Therefore even early positive results from online standardisation must be treated with caution; it is possible they are successful because of the groundwork laid in previous years' face-to-face meetings. If standardisation becomes a wholly online affair and examiners are no longer afforded face-to-face meetings, there is a risk that, over time, shared understanding will reduce and the virtual teams will no longer operate so successfully as a community of practice. It is equally possible, as noted earlier, that, over time, online standardisation processes will produce improved effects, as examiners become more familiar with new processes and procedures.

4 Discussion

This report has reviewed marker standardisation, mark schemes and their impact on marking quality; particularly in an environment of rapid introduction of technologies. Several of the research studies quoted in this review purport to show that online standardisation/marker training need not have an adverse impact on quality of marking, if properly implemented. This view is reinforced by the brief review of research into virtual meetings, which found that virtual meetings can work well if they are task focused, use appropriate technology, and provide appropriate training and support.

In addition to the findings that marking should at least get no worse if standardisation moves online, there is pretty clear evidence of substantial cost and time gains – again this is consistent between the online marker training and the online meeting/communication literature. Ofqual has statutory duties relating to the efficiency of qualifications provision, and awarding organisations will have clear motivations for adopting the most cost-effective solutions to major outlay activities. As such, arguments against adopting a more effective solution should be clear-cut and based on sound evidence and reasoning.

We know that there exist at the moment many questions that the online standardisation/marker training literature has not yet answered. For example, there is the question of longitudinal effects. It is possible that early positive results of online standardisation are 'riding on the coat-tails' of pre-existing quality assessment practice, but that over time the shared understanding and community of practice effects will be eroded, to the detriment of marking reliability. However, in contrast, maybe initial, negative reactions to new ways of working will be overcome in time and online standardisation will lead to higher quality in the medium to long-term. Perhaps a very important strand of research in this context pertains to marker retention. Tremain (2011, 2012) has started to survey this terrain, but it will be important to understand over time all aspects of the retention issue.

We saw in the findings section how many of the standardisation studies differed greatly from each other. Markers would be working to very different types of mark schemes, with different subjects, levels of responses, the markers themselves being very different. It is clear that any of these variables could affect outcomes profoundly.

It is also worth considering the nature of evidence that is available in this area. In the mark schemes area, there is much high quality theory-building research; thoughtful applications of first principles that are building up a body of knowledge and insights. In the online standardisation field, we see an increasing number of good quality research studies. These are, again, thoughtfully put together; as researchers carefully try to nail down variables, and to be explicit for those that they are not controlling[21]. Nonetheless, the great diversity of studies, and the limits on generalisability have been noted. From UK awarding organisation practice, we see the massive potential of item-level data; which is one of the most felicitous benefits of on-screen marking. And finally, from the online meetings literature, we can derive best practice for online meetings and provision of training.

There are many types of research study that need doing; more longitudinal studies, for example. But at a more strategic level, what is needed is integration of the theory building, the empirical studies, and use of the relatively recently available item-level data. Clear findings that follow from integrated, high quality research activities need to be brought into active use in live examinations. It is interesting that, as this body of research builds up, it should give insights into marker training and mark scheme development and use beyond those that were known about in the literature on 'mainstream' (i.e. non-online) standardisation.

[21] There is clearly a balance to be struck. As we saw in the mark schemes findings, an important action should be to investigate possible causes within mark schemes whenever an instance of marking inaccuracy is found. However, researchers must also take care to not chase after multiple idiosyncratic cases. A more strategic approach is needed.

Existing good practice in the field of mark schemes and marker training seems to have been agreed upon over time, and to work almost despite the fact that there are absences of hard evidence on many issues. Similarly, we have seen engineers build online meeting and communications systems, only for the researchers to follow in their wake and start to understand the implications of the new systems after their introduction. As new research evidence comes in to facilitate robust understanding of key issues, it needs to work within a framework of existing good practice, and high quality engineering principles. Research, good practice and sound engineering need to go hand-in-hand to produce high quality outcomes, rather than being in opposition to each other.

5 Methodology

5.1 Research questions

In respect of standardisation, the research was designed to address these issues:

- The impact that different forms of standardisation (i.e. online, face-to-face, webinars) have on marking reliability (and marker engagement).
- The impact that different stages of the standardisation process have on marking reliability.
- International practice for standardisation of examiners, how this takes place and any research showing the effectiveness of this. (Ofqual, 2013b, p. 7)

For the purpose of this review, the concept of standardisation is limited to the standardisation meeting as well as any relevant training undertaken prior to live marking; it does not extend to the on-going marking and moderation processes.

In respect of mark schemes, the research aimed to address the following issues:

- Do particular types of mark schemes work better with particular approaches to standardisation?
- How is a shared understanding of levels-based mark schemes facilitated via different approaches to standardisation?
- Are there distinct but subtle features of questions and their mark schemes that can affect accuracy?
- What is the relationship between different question/mark scheme types and marking strategies? How is this affected by marking mode (online or script-based)?
- Are there subject-specific effects?

Ofqual defined the research into online meetings in the following terms:

> … a brief review of any existing evidence relating to the effectiveness of online training/webinars compared to face-to-face training in other industries or for other types of training. This would be of contextual value rather than necessarily directly evidencing the effectiveness of online standardisation, and any conclusions drawn would need careful consideration in this respect.

The research questions pertaining to standardisation, mark schemes and online training each have their own sub-section of the findings section (pp. 4ff and pp. 17ff and pp. 26ff, respectively).

5.2 Method

Given the above assumptions, we can set out our method.

5.2.1 Search strategies

We employed a multi-pronged search strategy. It had the following elements:

- Keyword searches of well-known educational research databases: such as, Education Resources Information Center (ERIC), Digital Education Resource Archive (DERA), and so on.
- Hand searches of relevant journals.
- Keyword searches using specialist search engines: for example, Google Scholar.
- Hand searches of websites likely to have relevant reports in them – for instance, Ofqual's reliability compendium, Educational Testing Service (ETS)'s database of research reports.
- Contacts to known experts to seek out 'grey literature'. This included emails sent to colleagues and posts on groups in the professional social network Linked-in.

- Snowball searching: we found in Google Scholar any articles already identified for inclusion in the review, and checked any articles that had cited those articles using Google Scholar's 'cited by' function.

Full details of the implementation of this multi-faceted strategy are given at p. 46ff.

The strategy was used for the standardisation and mark schemes parts of the review. The third section on webinars, etc. was carried out in a less systematic, and more impressionistic manner. This was legitimate, given the broad nature of that topic, and its status as background for the main standardisation and mark scheme findings.

5.2.2 Selection of studies inclusion in the review

All potentially relevant studies located by the search strategies were listed in a spreadsheet. We then analysed the articles according to analytical criteria realised as columns. The columns, and their values were:

Column name	Range of values in column	Comment on column
In previous reviews	AQA NFER Neither	Where AQA = Meadows and Billington (2005) and NFER = Tisi et al (2013)
Aspect	Mark schemes Standardisation Both Both but mainly … Both tangentially, etc.	We coded the article to reflect its main focus amongst the themes of the review. However, mark schemes and standardisation are two interlinked topics and hence many articles mentioned both. Some interesting-looking articles only mentioned our topics tangentially – but we included them at least at the initial stage.
Impact on reliability	Yes or no	Because some articles wrote at length about mark schemes or standardisation, but not their impact on reliability.
Basis of paper	Review Empirical Argument	We mostly wanted to find articles with robust empirical findings, but also acknowledged that well-researched review articles could provide much information.
Quant, qual or mixed	Quantitative, qualitative or mixed	The main research method used in the study.
Analytical or data collection technique	Diverse entries	A more detailed description of the research methods used in the studies.
Relevance	High, medium or low	
Appropriateness of design	High, medium or low	
Appropriateness of analysis	High, medium or low	
Strength of findings	High, medium or low	
Total score	Number between 0 and 12	A sum of the four high, medium or low columns, where high = 3, medium = 2 and low = 1
Initial judgement – include?	Yes, no, maybe	An initial judgement by a coding researcher on whether or not to include the study in the review.

Table 6: Criteria used to analyse found articles

5.2.3 Analytical techniques employed

Literature reviewers must pay attention to the issue of evidential quality. Two researchers have put contrasting emphases on this matter. Wiliam (2003) writes about policy formation in UK assessment as follows:

> I think it is fair to ask whether we must we wait until all the evidence is in before things change. There is always the danger of making things worse, captured in the old adage that we cannot countenance change – things are bad enough as they are! The challenge for the

> educational research community is to provide policy-relevant findings when we cannot be certain about what to do. (Wiliam, 2003, p. 135)

Bennett (2011) critiques another article that Wiliam co-authored (Black & Wiliam, 2008). He argues that review articles can fall into a trap of overstating the effect of an educational innovation (formative assessment in this case) in order to argue for its introduction into policy and practice (Bennett, 2011, pp. 11 – 14)[22]:

> … the research does not appear to be as unequivocally supportive of formative assessment practice as it is sometimes made to sound. Given that fact, how might we improve the quality of the claims we make for the efficacy of formative assessment? An obvious first step should be exercising greater care in the evaluation of sources of evidence and in the attributions we make about them. Second, a clearer definition of what we mean by formative assessment … is essential to helping to abstract a class of things to study and make claims about. … Unless we understand the mechanisms responsible for change, we won't know if the effects are due to those mechanisms or to irrelevant factors. We also won't be able to predict the conditions, or population groups, for which the formative assessment is likely to work. (*ibid.*, at pp. 13 – 14)

In so far as Bennett's strictures apply to reviews of formative assessment, we think they also apply to reviews of standardisation and mark schemes.

There appears, at first glance, to be a broad gap between Wiliam's and Bennett's respective approaches. However, we try to straddle this chasm. We think that it remains a good thing to use reviews to inform policy, that one cannot wait until perfect evidence comes in (otherwise, one would get nothing done), but yet one should not – out of enthusiasm, naïve summarising or whatever cause – overstate effects in reviews. Black and Wiliam (1998, p. 43) use the term 'best evidence synthesis'. We think this is apt in this context[23].

In accordance with the discussion above, we seek to conduct 'best evidence synthesis'. Thus, we seek to draw out as many findings as we can and present them in a concise and clear manner for readers. However, in doing this, we also warn against over-interpretation; at points we draw out findings that have been surmised, even when the effects were not significant. We do this, because we feel it is better to highlight the evidence that is available, rather than wait for perfect evidence. However, readers should recall that tractable summaries of carefully written research reports often drop the various caveats and riders that were added to findings by responsible researcher-authors.

5.2.4 Terminological complexity

This report is written by United Kingdom (UK) researchers for a UK client. As such, we use terminology that is in current use here. So, for example, throughout the report we refer to markers and mark schemes. Many researchers working in other parts of the world would refer to the same concepts as 'raters', and 'rubrics' (or sometimes 'rating scales'), respectively. For consistency and coherence's sake, we stick to the UK terms, even where the original authors used the international/US variant.

In addition to the terms 'markers' and 'mark schemes' being problematic, we also need to draw attention to difficulties with the word 'reliability' in this context. Bramley (2007) is a verbally[24] precise and insightful article. It is cited in greater detail at p. 10, above. Essentially, Bramley (2007) warns us to only use the term reliability when context properly demands it. Bramley's strictures are

[22] In fact, Bennett's beef is not really with Black and Wiliam's original 2008 article, but rather with summary articles written by those authors, and/or subsequent researchers who have cited apparent effect sizes *sans caveats*.
[23] Newton (2007) is also informative on the evidential burdens on public bodies evaluating assessment systems.
[24] And numerically precise, of course!

borne out by the fact that some of the authors cited in this report (e.g. Wolfe, Matthews & Vickers, 2010) use validity coefficients where some might normally speak of reliability (also cited at p. 10, above). However, it would be inopportune to overlook important findings such as those of Wolfe, Matthews & Vickers (2010) just because they speak about validity indices, rather than reliability ones.

A wider notion is necessary. It is therefore useful that Newton (2005) has developed the notion of 'measurement inaccuracy'. He defines it as follows:

> Measurement inaccuracy is intended to encapsulate the variety of ways in which any set of assessment results will always depart from the mythical ideal of perfect accuracy for all students, due to the fundamental imprecision of educational assessment. This includes:
>
> - reliability deficit (e.g. where inaccuracy can be attributed to marker imprecision);
> - validity deficit (e.g. where inaccuracy can be attributed to test construction imprecision);
> - comparability deficit (e.g. where inaccuracy can be attributed to imprecision in the process of maintaining standards). (*ibid.*, at p. 420)

For Newton, it was important to emphasise the deficit, and hence the negative connotation of the word 'inaccuracy'. For us, this is less important, and thus we may speak of 'measurement accuracy'.

6 Data

By the writing-up phase of this project, the employment of the search and coding strategies described above led to the production of a spreadsheet table that listed 115 research reports. These 115 reports were then classified in terms of their relevance to the specification for this literature study (in the table below 'yes' means a report was definitely relevant to the project, whilst 'maybe' means that it might be). The relevance classification of these reports is shown in Table 7:

Aspect	Yes	Maybe	Row Total
Standardisation	19	6	25
Standardisation tangentially		3	3
Mark schemes	12	6	18
Mark schemes tangentially		5	5
Both	5	2	7
Both tangentially	2	9	11
Both, but mainly standardisation	3	2	5
Neither		2	2
Column Total	**41**	**35**	**76**

Table 7: Division of 'yes' and 'maybe' studies between aspects of research

The table appears to show that there are slightly more studies that discuss standardisation than discuss mark schemes. However, amongst the 18 studies that are coded as 'both', there is substantial discussion of mark schemes. The two papers that were coded as 'maybe' but which referred to 'neither' mark schemes nor standardisation are Tremain (2011, 2012). These papers are analyses of factors that affect the retention of markers and have been referred to in a discussion following the main findings in the standardisation section.

Table 8 shows the breakdown of studies by category, in terms of the type of research methods which they had used.

Aspect	Research method				Row Total
	Mixed	Qual	Quant	(blank)	
Standardisation	9	4	11	1	25
Standardisation tangentially	2			1	3
Mark schemes	4	5	9		18
Mark schemes tangentially	3	1	1		5
Both	2	3	2		7
Both tangentially	4	1	5	1	11
Both, but mainly standardisation	4		1		5
Neither	2				2
Column Total	**30**	**14**	**29**	**3**	**76**

Table 8: Research methods and study aspects in articles coded as 'yes' or 'maybe'

Overall, there is a predominance of quantitative studies within the standardisation literature particularly. In contrast to this, there is also a predominance of qualitative studies in the mark schemes literature. The basis for this, and some implications are discussed above, at pp. 23ff.

A few studies in Table 8 are not coded as to their research method. These are papers which were either reviews or arguments of principle. This includes Watts' (2007) mildly polemical argument about communities of practice, and their role in the world of online marking, which we refer to it in the relevant findings section.

7 Appendix 1: references[25]

Adie, L. E. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice, 20*(1), 91-106.

Adie, L. E., Klenowski, V. & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review, 64*(2), 223-240.

Ahmed, A. & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278.

Baird, J-A., Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, policy & practice, 11*(3), 331-348.

Baird, J-A., Hayes, M., Johnson, R., Johnson, S. & Lamprianou, I. (2013). *Marker effects and examination reliability: a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multi-level modelling*. http://ofqual.gov.uk/files/2013-01-21-marker-effects-and-examination-reliability.pdf.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2(1), 49-58. http://ehlt.flinders.edu.au/education/iej/articles/v2n1/barrett/barrett.pdf.

Baryla, E., Shelley, G. & Trainor, W. (2012). Transforming rubrics using factor analysis. *Practical Assessment, Research & Evaluation, 17*(4), 2.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Bird, F. L. & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme. *Assessment & Evaluation in Higher Education, 38*(5), 536-553.

Black, B., Suto, I. & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18*(3), 295-318.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-73.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters, 4*, 22-28.

Bramley, T. (2009). Mark scheme features associated with different levels of marker agreement, *Research Matters, 8*, 16-23.

Brindley, G. (1998). Assessment and reporting in language learning programs. *Language Testing 15*(1), 45–85.

Brown, N. (2009). New Zealand's National Education Monitoring Project: NEMP marking procedures and consistency. Paper BRO 99161, presented at the combined annual conference for 1999 of the *New Zealand Association for Research in Education and the Australian Association for Research in Education* Melbourne, Australia. November 29 - December 2nd, 1999. http://nemp.otago.ac.nz/PDFs/probe_studies/19brown.pdf.

[25] All web references were live on 16th September 2013.

Center for Educator Compensation and Reform (CECR). (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. http://cecr.ed.gov/pdfs/Inter_Rater.pdf.

Çetin, Y. (2011). Reliability of raters for writing assessment: analytic-holistic, analytic-analytic, holistic-holistic. *Mustafa Kemal University Journal of Social Sciences Institute*, 8(16), 471-486.

Chamberlain, S. & Taylor, R. (2010). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology, 42*(4), 665-675.

Chen, M., Lioub, Y., Wang, C-W., Fand, Y-W. & Chie, Y-P. J. (2007). TeamSpirit: design, implementation, and evaluation of a Web-based group Decision Support System. *Decision Support Systems, 43*, 1186–1202.

Chidambaram, L. & Jones, B. (1993). Impact of communication medium and computer support on group perceptions and performance: a comparison of face-to-face and dispersed meetings. *MIS Quarterly*, *17*(4), 465-491.

Cheung, K. M. A. & Chang, R. (2009). Investigating reliability and validity in rating scripts for standardisation purposes in onscreen marking. Paper presented at the *International Association for Educational Assessment (IAEA)* conference, Brisbane, Australia. http://www.iaea.info/papers.aspx?id=77.

Cohen, Y. (2007). Development and application of a simulation-based assessment center for non-cognitive attributes: screening of candidates to Tel Aviv University Medical School. Presented to *National Examinations Centre (NAEC) Tbilisi, Georgia*. 25th September 2007. www.naec.ge/images/doc/SXVA/SEM_Yoav-Cohen.pps.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation, 15*(3), 243-263.

Crisp, V. (2007). Researching the judgement processes involved in A-level marking. *Research Matters, 4*, 13-18.

Crisp, V. (2008a). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38*(2), 247-264.

Crisp, V. (2008b). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education, 79*(1), 1-12.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*(1), 1-21.

Donnellon, S., Gray, B. & Bougon, M. G. (1986). Communication, meaning, and organized action. *Administrative Science Quarterly*, *31*, 43–55.

Education Quality and Accountability Office (EQAO). (2012). *EQAO's technical report for the 2010 – 2011 assessments*. http://www.eqao.com/pdf_e/12/2011_TechnicalReport_en.pdf.

Elder, C., Barkhuizen, G., Knoch, U. & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elliott, V. (2013). Empathetic projections and affect reactions in examiners of 'A' level English and History. *Assessment in Education: Principles, Policy & Practice, 20*(3), 266-280.
Elwood, J. & Klenowski, V. (2002). Creating communities of shared practice: the challenges of assessment use in learning and teaching. *Assessment and Evaluation in Higher Education, 27*(3), 243-256.

Erickson, T., Kellogg, W. A., Shami, N. S. & Levine, D. (2010). Telepresence in virtual conferences: an empirical comparison of distance collaboration technologies. In: *Proceedings of CSCW 2010, 6-10*

February, Savannah, Georgia, USA. http://research.microsoft.com/en-us/events/nft2010/kellogg-virtualconferences.pdf.

Fahim, M. & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1), 1-16.

Gill, T. & Bramley, T. (2008). How accurate are examiners' judgments of script quality? In: *Proceedings of the British Educational Research Association (BERA) Annual Conference*, 3-6 September, Edinburgh.

Goodhue D. L. (1995). Understanding user evaluations of information systems. *Manage Science, 41*(12), 1827-1844.

Greatorex, J., Baird, J-A. & Bell, J. F. (2002). *'Tools for the trade': What makes GCSE marking reliable?* Paper presented at the conference *Learning Communities and Assessment Cultures: Connecting Research with Practice*. EARLI Special Interest Group on Assessment and Evaluation and the University of Northumbria. 28-30 August 2002, University of Northumbria UK.

Greatorex, J., & Bell, J. F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education, 23*(3), 333-355.

Haladyna, T. M. & Rodriguez, M. C. (2013) *Developing and validating test items*. New York: Routledge.

Hamilton, J. Reddel, S. & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System, 29*(4), 505–520.

He, Q. & Opposs, D. (Eds.) (2012). *Ofqual's reliability compendium*. Coventry: Office of Qualifications and Examinations Regulation.

Hinds, P. J. & Weisand, S. P. (2003). Knowledge sharing and shared understanding. In Gibson, C. B. and Cohen, S. G. (Eds.) *Virtual teams that work: creating conditions for virtual team effectiveness*. San Francisco, CA.: Jossey-Bass Business & Management.

Institute for Education Sciences/National Center for Education Statistics. (Undated). *NAEP item scoring process*. http://nces.ed.gov/nationsreportcard/contracts/item_score.asp.

International Baccalaureate Organisation (IBO). (Undated). *Theory of knowledge marking trial*, International Baccalaureate Organisation.

International Test Commission (ITC). (2013). *ITC guidelines on quality control in scoring, test analysis, and reporting of test scores*. 12th October, 2013, Version 1.2, Final Version. Document reference: ITC-G-QC-20131012. http://www.intestcom.org/upload/sitefiles/qcguidelines.pdf.

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: exploring variability. *Assessing Writing, 14*, 3-24.

Johnson, S., Johnson, R., Miller, L., & Boyle, A. (2013). Reliability of vocational assessment: an evaluation of level 3 electro-technical qualifications. http://ofqual.gov.uk/files/2013-01-17-c-and-g-reliability-of-vocational-assessment.pdf.

Jones, I., Swan, M. & Pollitt, A. (In press). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*.

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: attribute substitution in intuitive judgement. In Gilovich, T., Griffin, D. and Kahneman, D. (Eds.) *Heuristics and biases: the psychology of intuitive judgement*. Cambridge: Cambridge University Press.

Knoch, U. (2009). Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing 26*(2), 275-304.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. *Language Testing, 28*(2), 179-200.

Knoch, U., Read, J. & von Randow, J. (2007). Re-training writing raters online: how does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Lane, S. & Stone, C. A. (2006). Performance assessment. In Brennan, R. L. (Ed.), *Educational measurement* (4th edition). Westport, CT: ACE and Praeger Publishers.

Lantz, A. (2000). Meetings in a distributed group of experts comparing face-to-face, chat and collaborative virtual environments. *Behaviour & Information Technology, 20*(2), 111-117.

Leckie, G. & Baird, J-A. (2011). Rater effects on essay scoring: a multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

Lloyd, M. & Albion, P. (2005). Mistaking the tool for the outcome: using activity system theory to understand the complexity of teacher technophobia. In: *Proceedings of the Society for Information Technology and Teacher Education International Conference (SITE)*, 1-5 March, Phoenix, Arizona, USA.

Luppicini, R. (2007). Review of computer mediated communication research for education. *Instructional Science*, *35*, 141–185.

McAteer, E. & Harris, R. (2002). *Computer-mediated conferencing*. Bristol, UK: JISCinfoNet. http://strathprints.strath.ac.uk/3315/1/strathprints003315.pdf.

Meadows, M. & Billington, L. (2005). *A Review of the literature on marking reliability, report to NAA. National Assessment Agency*. https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf.

Meadows, M. & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: AQA Centre for Education Research and Policy. https://cerp.aqa.org.uk/research-library/effect-marker-background-training-quality-marking-gcse-english/how-to-cite.

National Institute for Testing and Evaluation (NITE) (Undated) *FAQs about the Psychometric Entrance Test*. https://www.nite.org.il/index.php/en/tests/psychometric/new-psych-faq.html.

Newton, P. E. (2005). The public understanding of measurement error. *British Education Research Journal, 31*(4), 419-442.

Newton, P. E. (2007). *Evaluating assessment systems*. London: Qualifications and Curriculum Authority.

O'Donovan, N. (2005). There are no wrong answers: an investigation into the assessment of candidates' responses to essay-based examinations. *Oxford Review of Education, 31*(3), 395-422.

Office of Qualifications and Examinations Regulation (Ofqual) (2012). *Corporate Plan 2012 – 2015*. http://ofqual.gov.uk/wp-content/uploads/2013/09/2012-05-15-corporate-plan.pdf.

Office of Qualifications and Examinations Regulation (Ofqual) (2013a). *Review of quality of marking in exams in A levels, GCSEs and other academic qualifications: interim report*. http://ofqual.gov.uk/files/2013-06-07-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-interim-report.pdf.

Office of Qualifications and Examinations Regulation (Ofqual) (2013c). *Office of Qualifications and Examinations Regulation (Ofqual): annual report and accounts 2012-13*. http://ofqual.gov.uk/documents/annual-report-and-accounts/.

Office of Qualifications and Examinations Regulation (Ofqual), Welsh government and Council for the Curriculum, Examinations & Assessment (CCEA) (2011). *GCSE, GCE, Principal Learning and Project Code of Practice: May 2011*. http://ofqual.gov.uk/files/2011-05-27-code-of-practice-2011.pdf.

Paul, D. & McDaniel, R. R. (2004). Effect of interpersonal trust on virtual collaborative relationship performance. *MIS Quarterly*, 28(2), 183-227.

Pell, G., Homer, M. S. & Roberts, T. E. (2008). Assessor training: its effects on criterion-based assessment in a medical context. *International Journal of Research & Method in Education, 31*(2), 143-154.

Pinot de Moira, A. (2011). *Effective discrimination in mark schemes*. Centre for Education Research and Policy. https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-APM-05042011.pdf.

Pinot de Moira, A. (2012). *Levels-based mark schemes and marking bias*. Centre for Education Research and Policy.

Pinot de Moira, A. (2013). *Features of a levels-based mark scheme and their effect on marking reliability*: Centre for Education Research and Policy.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157-170.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J. & Davidson, M. (2008). *Improving the quality of GCSE assessment*: Report commissioned by Qualifications and Curriculum Authority. http://www2.ofqual.gov.uk/files/Improving_the_Quality_of_GCSE_Assessment_final.pdf.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.

Royal-Dawson, L. & Baird, J. A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice, 28*(2), 2-8.

Shaw, S. (2002). The effect of standardisation training on rater judgement and inter-rater reliability for the revised CPE writing paper 2. *Research Notes*, *8*, 13-18.

Smart C. & Shiu J. (2010) *Report on a CE English Language marker re-training programme*: (December 2009- June 2010). Unpublished report on Hong Kong Certificate of Education Examination (HKCEE) English Language examination.

Stahl, G., Koschmann, T. & Suthers, D. (2006). Computer-supported collaborative learning: an historical perspective. In Sawyer, R. K. (Ed.), *Cambridge handbook of the learning sciences.* Cambridge, UK: Cambridge University Press.

Stobart, G. (1998). *Key Stage 3 English: Review of External Marking in 1997*. Unpublished report to School Curriculum and Assessment Authority (SCAA).

Suduc, A. M., Bizoi, M. & Filip, F. G. (2009). Exploring multimedia web conferencing, *Informatica Economica, 13*(3), 5-17.

Suto, W. I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*(2), 213-233.

Suto, W. I. & Nádas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24*(3), 335-377.

Suto, W. I., Greatorex, J. & Nádas, R. (2009) Thinking about making the right mark: using cognitive strategy research to explore examiner training, *Research matters, 8,* 23-32.

Sweiry, E. (2012). Conceptualising and minimising marking demand in selected and constructed response test questions. *Paper presented at the Association for Educational Assessment Europe Annual Conference*, Berlin.

Texas Education Agency Student Assessment Division (TEASAD). (2013a). *2013-2014 holistic rating training requirements*. http://tinyurl.com/oeposx5.

Texas Education Agency Student Assessment Division (TEASAD). (2013b). *TELPAS Texas English Language Proficiency Assessment System: manual for raters and test administrators grades K–12.* http://tinyurl.com/q9za7tl.

Thompson, L. F. & Coovert, M. D. (2003). Teamwork online: the effects of computer conferencing on perceived confusion, satisfaction and post-discussion accuracy. *Group Dynamics, 7*(2), 135–151.

Tisi, J., Whitehouse, G., Maughan, S. & Burdett, N. (2013). *A review of literature on marking reliability research*. http://ofqual.gov.uk/files/2013-06-07-nfer-a-review-of-literature-on-marking-reliability.pdf.

Tremain, K. (2011). *Carry on examining: what predicts examiners' intentions to continue examining?* Centre for Education Research and Policy. https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-KMT-16112011_0.pdf.

Tremain, K. (2012). *Carry on examining: further investigation*. Centre for Education Research and Policy. https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_MO_KMT_01062012_0.pdf.

Turban, Liang, T-P. & Wu, S. P. J. (2011). A framework for adopting collaboration 2.0 tools for virtual group decision making. *Group Decision and Negotiation*, 20(2), 137-154.

Vailaitis, R., Akhtar-Danesh, N. Eva, K., Levinson, A. & Wainman, B. (2007). Pragmatists, positive communicators, and shy enthusiasts: three viewpoints on web conferencing in health sciences education, *Journal of Medical Internet Research, 9*(5), 2007.

Veerman, A. L., Andriessen, .J. E. B. & Kanselaar, G. (1999). Collaborative learning through computer-mediated argumentation. *Paper presented at the Conference on Computer Supported Collaborative Learning (CSCL 99)*, San Francisco, California.

Vickers, D. & Nichols, P. (2005). The comparability of online vs. stand-up training. *Paper presented at the 35th National Conference on Large-Scale Assessment*, San Antonio, TX.

Warkentin, M., Sayeed, L. & Hightower, R. (1997). Virtual teams versus face-to-face teams: an exploratory study of a web-based conference system. *Decision Science, 28*(4), 975-996.

Way, W. D., Vickers, D. & Nichols, P. (2008). Effects of different training and scoring approaches on human constructed response scoring. Paper presented at the *Annual meeting of the National Council on Measurement in Education*, April, New York City.

Wolfe, E. W. & McVay, A. (2010). Rater effects as a function of rater training context. http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf.

Wolfe, E. W. & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37.

Wolfe, E. W., Matthews, S. & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment, 10*(1).

Wood, W., Mueller, J., Willoughby, T., Specht, J. & Deyoung, T. (2005). Teachers' perceptions: barriers and supports to using technology in the classroom. *Education, Communication & Information*, *5*(2), 183-206.

Wiliam, D. (2003). National curriculum assessment: how to make it better. *Research Papers in Education*, 18(2), 129–137.

Xi, X. & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.

8 Appendix 2: implementation of search strategy

This appendix indicates the search strategies employed to locate relevant literature for the review. The searches involved the use of key words for the following categories of sources:

1. database searches
2. hand searches of journals
3. web-site searches, including use of specialist search engines

The keywords were selected to represent the concepts most relevant to the focus of the report but contained sufficient breadth to capture a range of documents from different disciplines bearing in mind the variations in terminology across national and international journals. Searches were restricted to the years 2005 to 2013.

8.1 Database searches

A brief description of the databases searched and the keywords is provided below. For the main databases the keywords were initially used systematically, in the combinations indicated in the tables. The search results from these key word combinations were tracked in terms of the number of new and relevant documents generated so a more concise set of keywords could be used subsequently. The search strategy was also sensitive to the size and particular focus of each database.

The document lists generated by the searches were sorted from the most to the least relevant where this data base facility was available. The lists were then inspected for duplication and abstracts hand-searched for relevance before being selected for closer examination.

8.1.1 British Education Index (BEI)

The British Education Index provides information on research, policy and practice in education and training in the UK. Sources include education and training journals, mostly published in the UK, plus books, reports, series and conference papers. The database covers all aspects of education from preschool to higher education from 1975 to date.

Accessed via the University of Nottingham, August 7th 2013.

8.1.2 Education Resources Information Center (ERIC)

The ERIC (Educational Resources Information Center) database is sponsored by the US Department of Education to provide extensive access to education-related literature.

Accessed via the University of Nottingham, August 7th 2013.

KEYWORD 1	KEYWORD 2	BEI			ERIC		
		Number generated by search	Number relevant	No of new and relevant items	Number generated by search	Number relevant	No of new and relevant items
examiner	training	19	5	5	49	11	11
examiner	standardisation	1	1	0	6	1	0
examiner	co-ordination	0	0	0	2	0	0
examiner	meeting	2	1	1	15	5	4
examiner	mark scheme	4	2	1	6	4	1
examiner	marking scheme	3	2	0	5	4	1
examiner	marking protocols	11	10	3	7	7	1
examiner	marking rubrics	0	0	0	1	0	0
examiner	marking instructions	1	1	0	4	2	0
marker	training	3	1	0	45	6	1
marker	standardisation	2	1	0	8	1	0
marker	co-ordination	1	0	0	12	0	0
marker	meeting	2	1	0	20	2	0
marker	mark scheme	3	2	0	4	2	0
marker	marking scheme	3	2	0	4	2	0
marker	marking protocols	0	0	0	2	2	0
marker	marking rubrics	2	1	0	3	3	2
marker	marking instructions	0	0	0	10	2	1
marking	training	17	5	2	52	12	2
marking	standardisation	4	2	1	4	2	0
marking	co-ordination	0	0	0	0	0	0
marking	meeting	14	5	4	9	4	0
marking	mark scheme	6	2	0	9	4	1
marking	marking scheme	16	3	1	22	7	2
marking	marking protocols	17	10	0	13	8	1
marking	marking rubrics	5	1	0	11	3	0
marking	marking instructions	5	1	0	120	4	0
assessor	training	11	1	0	43	2	2
assessor	standardisation	0	0	0	0	0	0
assessor	co-ordination	0	0	0	0	0	0
assessor	meeting	1	1	1	4	0	0
assessor	mark scheme	0	0	0	1	0	0
assessor	marking scheme	1	0	0	2	0	0
assessor	marking protocols	2	0	0	1	0	0
assessor	marking rubrics	0	0	0	1	1	0
assessor	marking instructions	0	0	0	1	0	0
standardisation	meeting	4	0	0	17	2	1
standardisation	examinations	4	2	0	37	1	0

8.1.3 Scopus (Elsevier)

Scopus is the largest abstract and citation database of peer-reviewed literature and delivers a comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences and Arts & Humanities.

Accessed via the University of Nottingham, August 8th 2013.

8.1.4 Applied Social Sciences Index and Abstracts (ASSIA)

The Applied Social Sciences Index and Abstracts on the Web is an indexing and abstracting tool covering health, social services, psychology, sociology, economics, politics, race relations and education. ASSIA provides a comprehensive source of social science and health information for the practical and academic professional from 16 countries including the UK and US.

Accessed via the University of Nottingham, August 8th 2013.

KEYWORD 1	KEYWORD 2	SCOPUS			ASSIA		
		Number generated by search	Number relevant	No of new and relevant items	Number generated by search	Number relevant	No of new and relevant items
examiner	training	133	8	8	34	0	0
examiner	standardisation	3	1	0	3	0	0
examiner	co-ordination	6	0	0	2	0	0
examiner	meeting	21	2	0	5	0	0
examiner	mark scheme	6	4	0	0	0	0
examiner	marking scheme	7	4	0	0	0	0
examiner	marking protocols	5	5	0	0	0	0
examiner	marking rubrics	1	0	0	0	0	0
examiner	marking instructions	0	0	0	0	0	0

8.1.5 Education-line

This search interface gives access to BEI's Educati*on-line* collection which contains mostly, but not exclusively, conference papers, presented to the BEI by their authors.

Accessed via the University of Nottingham, August 8th 2013.

KEYWORD	Education-line		
	Number generated by search	Number relevant	No of new and relevant items
examiner	2	1	1
standardisation	6	0	0
marker	2	1	0
marking	17	4	2
assessor	3	1	0

8.1.6 Centre for the Economics of Education (CEE)

The CEE is a multidisciplinary centre with three partners: The Centre for Economic Performance at LSE; the Institute for Fiscal Studies; and the Institute of Education. The CEE seeks to undertake systematic and innovative research in the field of the economics of education by applying the latest techniques of empirical analysis.

Accessed: August 8th 2013 at http://cee.lse.ac.uk

KEYWORD	CEE		
	Number generated by search	Number relevant	No of new and relevant items
examiner	1	0	0
standardisation	0	0	0
marking	0	0	0
mark scheme	0	0	0
rubrics	0	0	0

8.1.7 Digital Education Resource Archive (DERA)

The IOE UK Digital Education Repository Archive (DERA) is a digital archive of documents published electronically by government and other relevant bodies in the areas of education, training, children and families.

Accessed: August 8th 2013 at http://dera.ioe.ac.uk

KEYWORD 1	KEYWORD 2	DERA		
		Number generated by search	Number relevant	No of new and relevant items
standardisation	meetings	564	4	3
standardisation meetings	examiners	1	1	1
standardisation	marking examinations	2	0	0

8.1.8 ETS ReSEARCHER Database

ETS ReSEARCHER is a database that contains information on ETS-authored or published works, such as ETS Research Reports, ETS Research Memorandums, or publications written by ETS researchers and published by third parties, such as scholarly journals.

Accessed: 9th August 2013 at http://search.ets.org/researcher/

KEYWORD	ETS ReSEARCHER		
	Number generated by search	Number relevant	No of new and relevant items
Rubric	17		
Scoring constructed response	500		
Rater standardize	302		
Rater cognition	178		
Standardization meeting	156		
Rater training	402		

8.2 Hand searches

Each journal has a different focus and use of the key words was adapted to suit the style, focus and terminology of the journal. However, where the electronic facilities were available groups of journals were searched together. Some searches generated large numbers of results which were sorted automatically by relevance and then the top 100 most relevant results were inspected by hand.

8.2.1 Hand-searching of journals published by Taylor and Francis

The following journals were searched using the Taylor and Francis platform.

Accessed via the University of Nottingham, August 12th 2013.

- Applied Measurement in Education
- Assessment and Evaluation in Higher Education
- Assessment in Education: Principles, Policy & Practice
- Educational Assessment
- Journal of Vocational Education and Training
- Measurement: Interdisciplinary Research and Perspectives

(12 new items identified)

8.2.2 Hand searches of US journals

Accessed: Friday, 09 August 2013

- Practical Assessment, Research & Evaluation
- Measurement: Interdisciplinary Research and Perspectives
- International Journal of Testing
- Educational Measurement: Issues and Practice
- Educational Assessment
- Applied Measurement in Education

(12 new items identified)

8.2.3 Hand-searching of individual journals

Accessed August 12th 2013

- Applied Psychological Measurement
- International Journal of Selection and Assessment
- Educational and Psychological Measurement

(No new items identified)

- Cambridge Assessment: Research Matters

(10 new items identified)

8.3 Website searches

Use of the key words was adapted to suit the style, focus and terminology of each site accessed.

8.3.1 Ofqual reliability compendium search

Accessed: Thursday, 8th August 2013

http://ofqual.gov.uk/standards/research/reliability/compendium/

(14 new items identified)

8.3.2 Institute of Education (IoE) Library and Archives

Accessed: Friday, 9th August 2013 using Institute of Education Library Catalogue

Search of electronic resources only.

(48 new items identified but largely National Curriculum test mark schemes)

8.4 Contacts approached for grey literature

8.4.1 Grey literature contacts (UK)

Names have been removed to protect colleagues' confidentiality.

Organisation	They responded (Y/N)	We responded to them (e.g. said thank you) (Y/N)	Their response/suggestion has been followed up (Y, N, N/A)
AQA	N		
Cambridge Assessment	Y	Y	N/A
Pearson	N		
City & Guilds	N		
Royal College of Surgeons	Y	Y	Y
NFER	Y	Y	N/A
IoE	Y	Y	Y
Cambridge Exams Limited	N		
Cambridge English Language Assessment	N		
Standards & Testing Agency	Y	Y	Y
CEM centre	N		
Was IoE	N		
KCL	Y	Y	N

8.4.2 Grey literature contacts (Overseas)

Organisation	Country	They responded (Y/N)	We responded to them (e.g. said thank you)	Their response/suggestion has been followed up (Y, N, N/A)
CITO	NL	Out of Office	N/A	
ETS	USA	Y	Y	Y
Uni of Oslo	Nor	N		
	Estonia	Y	Y	Y
	DK	Y	Y	Y
	USA	Y	Y	Y
Cito	NL	N		
NITE	IL	N		
ACER	Aus	N		
Uni of Umea	Swe	Y	Y	N/A
HKEAA	Hong Kong	Y	Y	Y

8.4.2.1 Linked-in groups

A request for assistance was posted on groups in the professional networking site, Linked-in[26].

IAEA: http://tinyurl.com/k4uhqrc

[26] It is necessary to be a member of Linked-In (and possibly the groups) to see these posts.

AEA-E: http://tinyurl.com/nypvxny

The requests generated a range of suggestions from colleagues in the United Kingdom, Canada, Singapore and the Maldives, amongst others. Any suggested references from such sources were checked out.

9 Appendix 3: background on standardisation

9.1 UK code of practice

In order to portray standardisation's implementation in general qualifications in England, we summarise from the GCSE, GCE, Principal Learning and Project Code of Practice (CoP) (Ofqual et al, 2011). Standardisation under the CoP is mandatory (*ibid.* at p. 21), and the standardisation section of the Code covers the following issues:

- training, monitoring and supervising examiners
- checking the work of examiners
- action to be taken if marking instructions are not followed
- reviewing examiner performance. (*ibid.*)

The standardisation provisions in the CoP apply equally to traditional and online marking (*ibid.*).

Standardisation is a hierarchical concept; the aim of the procedure being to communicate the awarding organisation's concept of the 'true' standard embodied in a mark scheme to a succession of more junior staff (referred to as: chief examiners, principal examiners, assistant principal examiners, team leaders and examiners). This concept may also be referred to as a 'cascade' approach. It can be contrasted with a consensual approach, in which a group of professionals (markers and supervisors) work together to arrive at a joint appreciation of the standard embodied in a mark scheme (cf. Baird, Greatorex & Bell, 2004; Adie, 2013).

The CoP envisages that training for markers might vary given their diverse experience – for example, differentiated training could be provided for: first-time markers, markers new to the particular awarding organisation, and markers new to the particular unit or component (*ibid.* at p. 24). The CoP also envisages the mentoring for junior markers (*ibid.*).

All markers must have studied the mark scheme and marked a provisional sample of candidate work before standardisation (*ibid.*). The standardisation process is required to contain the following elements:

i. an administrative briefing from the awarding organisation that includes reference to this section of the Code, awarding organisation procedures, time schedules, administrative documentation and contact points
ii. an explanation from the principal examiner of the nature and significance of the standardisation process
iii. a briefing from the principal examiner on relevant points arising from current examinations, drawing as necessary on relevant points made about previous examinations in chief examiners' reports and regulatory monitoring reports
iv. a discussion of marking issues, including:
    - full consideration of the mark scheme in the context of achieving a clear and common understanding of the range of acceptable responses and the marks appropriate for each item being marked, and comparable marking standards for optional questions
    - handling of unexpected, yet acceptable, answers
v. the marking of a number of common, clean responses sufficient to:
    - illustrate the range of performance likely to be demonstrated by the candidates in an examiner's allocation

- help consolidate a common understanding of the mark scheme, including any criteria for the assessment of written communication (*ibid.* at p. 25)

It is interesting, given the concerns that have given rise to this project, that the CoP does not explicitly mandate the delivery mode for the elements cited above. It does not say, for instance, that they must be delivered by face-to-face meeting[27]. But neither does it acknowledge explicitly that remote (e-facilitated) standardisation is permissible.

Following the standardisation process outlined above, markers must mark a sample of at least 10 items of the type they have been allocated, and demonstrate sufficient marking care, accuracy and consistency before they are cleared to carry out live marking (*ibid.* at p. 26). Audit trails must be maintained throughout live marking (*ibid.*), and there must be periodic checks of marking. Such checks are defined differently for traditionally-marked scripts and online marking (*ibid.*, at pp. 27 – 28). Stipulations are in place for removing markers and for adjusting any erroneous marks given by an aberrant marker (*ibid.*, at p. 28).

9.2 Operational practice in other locations worldwide

In order to provide context for UK practice in GCSEs and A levels as exemplified by CoP stipulations, we have collected examples of rater training/standardisation in other worldwide locations. We contacted known experts in 11 jurisdictions around the world. We also posted requests for information on professional forums held within a social network website. This request engendered replies from Canada, Singapore and the Maldives, amongst others. (See pp. 52ff, above for details of our 'grey literature' searches.) Additionally, we conducted top-up searches to find further descriptions of worldwide standardisation practice.

Some sources give a general picture of practice in a particular country – often re-assuring general readers of the existence of marker training (for example – on Israel's Psychometric Entry Test (NITE, Undated)). Other sources locate standardisation as an element (sometimes a small element) within a wider argument concerning the quality of marking on a particular assessment. Such sources speak of the multiple marking which is perceived to be a guarantor of quality, mark scheme/rating scale development and statistical analyses, amongst other things. Cheung & Chang's (2009) paper on an examination in Hong Kong is a good example of this phenomenon. It describes how scripts for standardisation were drawn from a stratified random sample (*ibid.* at p. 3). It also shows how Rasch analysis was used to derive a 'fair average' score for standardisation scripts, based on the judgement of experienced markers (*ibid.*). But it also describes findings of Facets analysis to investigate various features of marker performance, correlational analysis and even statistical analysis of the sentence complexity of scripts (*ibid.* at p. 8). In this way, we see researchers and examiners using standardisation as part of a wider process of quality assurance in marking, rather than in isolation.

Other descriptions of marker training give more details. For example, IES/NCES (Undated) give a general description of item scoring practices, but – within that – give more detail about marker recruitment and training on the National Assessment of Educational Progress (NAEP) in the USA. 'NAEP scorers' are carefully screened as part of their recruitment process, they participate in a training process. This process includes:

- extensively reviewing the scoring guides
- discussing the anchor papers
- having scorers score and discuss sets of practice papers (*ibid.*)

[27] These are generally referred to as either 'standardisation' or 'co-ordination' meetings in the UK.

Following training, scorers must pass a qualification test before live marking. The qualification test involves:

> … items that are identified by test developers and scoring directors as particularly challenging to score. Each scorer must score a set of papers that has been pre-scored by NAEP content and scoring experts. If the scorer does not have a high enough level of agreement with the pre-assigned scores (70 per cent or more), he or she is not allowed to score that item. (*ibid.*)

Scrutiny of markers continues throughout the marking process, with markers having 'calibration papers' seeded throughout the process.

Cohen (2007) outlines the marker training day conducted in a new assessment in an Israeli medical school. The training had the following elements:

> **Train the raters workshop**
>
> Half a day – mandatory for participation
>
> Groups of 20 faculty [staff] [in] each [session]
>
> Include:
>
> - Overview of new admission process
> - Awareness of biases (halo, cultural, etc.)
> - Introduction of rating scales and behavioral anchors
> - Actual rating exercises based on videos of 'standardized' candidates prepared in advance
> - Calibration of raters through open discussion of metrics and reference to group ratings

Summaries of practice in three locations are set out at pp. 60ff. The sources of the information vary; the Ontarian practice (pp. 62ff) is extracted from a technical manual, whilst information about rater training in the Texas English language assessment system (TEASAD, 2013b, pp. 65ff) is aimed at teachers. Hence the latter is somewhat less technical than the former.

All of the three training approaches – either explicitly or implicitly – set out to ensure marking accuracy. However, being operational procedures, rather than research articles, none of the documents report values on reliability, validity or other relevant coefficients.

All three operational processes exemplify how the ensuring of marking accuracy is a detailed process with many phases. The Ontarian procedures show how those jurisdictions use field testing as a way to prepare robust mark schemes prior to operational use. The Ontarian process shows their standardisation/training to be a hierarchical process with separate training for leaders and scoring supervisors on the one hand, and 'ordinary' markers on the other. Texas has (initial) training for novice raters and re-calibration for experienced colleagues (TEASAD, 2013a).

The Ontarian processes appear to involve face-to-face training; although the Ontarian markers are apparently uploading data using PDAs; thus they may be getting a benefit of online standardisation while retaining the social elements of standardisation that participants sometimes appear to value. The Texas training programmes are online. The Texan authorities emphasise the benefits of this (flexibility, ability to work at home, etc.) to their raters.

In contrast to the hierarchical approaches to standardisation that appear to predominate in recent literature, Brown (1999) describes a consensual approach to standardisation under New Zealand's National Education Monitoring Project (NEMP). NEMP was a national sampling and monitoring assessment carried out in years four and eight of the NZ system. Children undertook tasks – either individually with a teacher, or in groups. These tasks were video recorded. Brown (1999) describes

a process she calls 'cross marking'. In this process teacher-markers viewed a succession of video performances and discussed proper scoring in a group of up to 20. This process was repeated until consensus was felt to have been reached on features of performances that were associated with particular scoring levels.

Brown (2009, p. 10) argues that cross marking enhances validity as follows:

> Cross-marking allows markers to apply their professional judgement to these issues and then receive feedback from others. In doing so, markers develop a robust understanding of the task construct and the qualities associated with each grade, which can then be applied to the range of responses that are generated in the NEMP data. Cross-marking therefore facilitates the development of a sense of 'ownership' amongst markers which is used to aid consistency when making judgements on student performance. Discussions which occur during cross-marking also allow markers to share their experience of a range of student responses, and in so doing they may collectively identify the need for additional categories which are not covered by the existing marking criteria. Cross marking therefore enhances the validity of the marking process by allowing a more accurate and representative picture of student achievement to emerge.

There are two examples of guidelines or standards which provide insight into what is considered good practice in standardisation around the world. ITC (2013) is a set of guidelines for quality control in scoring, test analysis, and reporting of test scores.

The Guidelines' provisions pertaining to rating performance tests, work samples, role plays, interviews, etc. include the following:

> 2.3.3.1. Make sure that performance on tests, work samples, role playing, and interviews are rated by trained assessors who have the requisite knowledge and experience, as well as credentials, training or appropriate formal education.
>
> 2.3.3.2. The instructions for rating open-ended (OE) responses should be clear and well-structured. A pre-test of the OE should take place to help constructing the instructions.
>
> 2.3.3.3. Use range-finding activities to pinpoint examples of OE student's responses at each rubric point. Involve sample papers in scoring training activities.
>
> 2.3.3.4. Require raters to participate in training sessions before undertaking the rating process. Training enables them to become familiar with rating instructions and to practice the scoring of assessment material before they become authorized to evaluate actual test taker responses.
>
> 2.3.3.5. Assess raters' competence based on their training, prior to having them undertake operational rating.
>
> 2.3.3.6. Try to use at least two raters for each assessment, depending on costs and availability of resources.
>
> 2.3.3.7. When there is only one rater for all test takers (due to financial or other considerations) use two raters per sample (e.g., 10% of the data) to estimate scoring reliability, depending on stakes, length of test, and other factors.
>
> 2.3.3.8. If computer scoring of OE items is used, ensure that the scoring is monitored by a human rater. Justify the use of computerized scoring on the basis of research before using it operationally.
>
> 2.3.3.9. Ensure that raters work independently of one another.
>
> 2.3.3.10. Apply statistical procedures to assess the reliability of the rating process, i.e., by computing appropriate measures of inter-rater agreement as well as differences between

> raters within and across raters by checking the degree of correspondence as well as the differences between raters and using appropriate measures to eliminate correlation coefficients between rater results that are similar just by chance.
>
> 2.3.3.11. Monitor the rating quality periodically in real time, so feedback will be available.
>
> 2.3.3.12. If a rater is not meeting expectations, (ratings are unreliable or not close enough to those of other raters) inform the person accordingly and consider retraining; do not hesitate to replace the person if the problem is not resolved.
>
> 2.3.3.13. Develop policies for dealing with large discrepancies between raters. When differences are small, they should be averaged or summed to avoid rounding problems. When there are large discrepancies, an experienced rater may mediate to resolve them. (*ibid.* at pp. 18 – 19)

The frame-of-reference training outline (pp. 60ff) from the US Center for Educator Compensation Reform (CECR) is in fact a set of recommendations for how 'frame-of-reference' training (cf. communities of practice, above at p. 9) could be carried out, rather than an operational manual. However, the outline reads like a set of standards, or an extract from a code of practice; as such it is included in Appendix 3.

The CECR recommends making raters aware of common errors, such as: similarity, leniency, halo effect, central tendency, inconsistency and context effects (see p. 60, below). It is worth comparing such stipulations, however, with the findings of researchers such as Knoch (2011), who had mixed success when implementing a similar scheme.

Whilst the CECR's recommendation to make raters aware of potential biases may be too idealistic to work in practice, its insistence on permitting those professionals to 'see the big picture' appears to be well founded. Furthermore, the following passage, which concludes the framework, is surely apposite for all marking processes:

> Even detailed rubrics, trained raters, and good evidence will not make performance assessment a completely objective process. Some professional judgment will always be called for in assessing performance in professional jobs.
>
> The goal of rater training is not to eliminate professional judgment but to guide and focus it. (CECR, 2012, p. 27)

9.3 General evidence on standardisation

There is a well-established body of research on standardisation[28]. This is summarised well in several places, including: Meadows and Billington's review (2005, pp. 50 – 52), Haladyna and Rodriguez's recently updated textbook (2013, pp. 254 – 255), and Lane and Stone's chapter on performance assessment in the latest edition of the professional manual 'Educational Measurement' (2006, pp. 400 – 401).

Much of the collated research evidence on standardisation/rater training accords with what has been seen from the UK and international practice summarised above. Rater training is 'one of the most important tools system administrators have to improve agreement' (CECR, 2012, p. 15); however, it cannot remove measurement inaccuracy completely, and will not work in isolation from other quality assurance mechanisms (Haladyna & Rodriguez, 2013, p. 255).

The stages of standardisation/rater training outlined in the summaries of practice above have also been investigated in the research literature. Pre-marking qualification (or credentialing – to use the Americanism) is recommended (*ibid.*, at p. 254). The research evidence on different aspects of

[28] In fact, since much of the research is US in origin, it tends to refer to 'rater training' rather than standardisation.

training is not clear, however. Some of the prescriptions, such as monitoring the quality of training, and providing standardised training packages to different marking/rating locations (CECR, 2012, p. 15) seem somewhat self-evident. Others, such as the desirability of making raters aware of different sources of measurement inaccuracy, the duration of training or the approach – hierarchical as opposed to consensual – produce mixed results.

Barrett (2001) provides the following useful summary, which sets out what training can achieve and how one can determine whether a marker/rater has in fact been well trained:

> Training is a necessary condition if rater inconsistencies are to be minimised, if not eliminated. Mills, Melican and Ahluwalia (1991) argue that training of raters should achieve four important outcomes. First, training provides a context within which the rating process occurs. Second, training defines the tasks to be performed by the raters. Third, training minimises the effects of variables other than item difficulty from the rating process. Fourth, training develops a common definition of the minimally competent candidate. Furthermore, there are three measurable criteria that can be used to determine whether a rater is well trained (Reid, 1991). First, ratings should be stable throughout the rating process. Second, ratings should reflect the relative difficulties of the test items. Third, ratings should reflect realistic expectations of the expected performance of the candidates. However, the big question remains, how should raters be trained? Hambleton and Powell (1983) argue that this is a difficult question to answer due to the poor documentation of training procedures in most of the reports of standard setting studies. (*ibid.* at p. 51)

Two of the more important papers cited by Meadows and Billington (2005) are Shaw (2002) and Baird, Greatorex and Bell (2004). These are now summarised in turn. Shaw (2002) investigated a standardisation process with multiple iterations. In general, inter-rater reliability benefited from the training iterations, however, findings were not straightforward. The two later iterations produced 'see-saw' and 'erratic' results (*ibid.*, at p. 16). Whilst inter-rater reliability could improve, the numbers of markers whose severity was off track (too severe or too lenient) changed over time. The bar charts for iterations four and five are almost mirror images of each other; seemingly, unduly harsh markers in iteration four may have over-compensated in iteration five and thus become unduly lenient.

Baird, Greatorex and Bell (2004) investigated the measurement accuracy of marking under three standardisation conditions: a consensual meeting, a hierarchical meeting and no meeting at all. These three approaches produced results with very similar levels of error amongst the three conditions; indeed, generalisability coefficients were identical to two significant figures (*ibid,* at p. 343). These results were surprising to the authors (*ibid.*, at p. 345). The paper is very important in the context of communities of practice, and is discussed alongside other studies at pp. 9ff, above. The 'surprising' results perhaps also speak of a cadre of markers whose skills and professional attitudes were well-established and which endure (can't be artificially 'forgotten') when the markers participate in synthetic research exercises (cf. also Chamberlain & Taylor, 2010, discussed above). However, both the Baird, Greatorex and Bell (2004) and the Shaw (2002) studies re-enforce the observation made above that we appear to know that certain kinds of standardisation work, but not (yet) why or how. That this is true of 'conventional' standardisation is a point worth recalling as we begin to look at novel forms of standardisation.

9.4 Examples of standardisation practice around the world

9.4.1 US Center for Educator Compensation Reform (CECR) recommendations

The CECR states its remit as follows:

The primary purpose of CECR is to support Teacher Incentive Fund (TIF) grantees in their implementation efforts through provision of sustained technical assistance and development and dissemination of timely resources. (CECR, 2012)

In an appendix to its document on inter-rater agreement of teacher and principal ratings, it sets out an outline of 'frame-of-reference training'. This is reproduced below:

9.4.1.1 Frame-of-reference training outline

1. Provide a process overview to give the observers the big picture.

- Purpose of observations.
- Frequency and length of observations.
- Use of pre- or post-conferences, collection of artifacts.
- How results will be used.
    - Feedback to person being evaluated.
    - Coaching/assistance for performance improvement.
    - Goal setting.
    - Administrative consequences for good and poor performance.

2. Explain the rating dimensions (standards of performance & rubrics).

- Review rubrics.
- Explain how rubrics are consistent with or represent organization's vision of good practice.
- Discuss questions about concepts or wording.

3. Help raters identify and put aside their own biases.

- All observers bring beliefs about what good teaching looks like, which can influence what they see and how they evaluate it.
- Explain that observers need to be able to separate these beliefs from the observation, especially when observing a different style, level, or subject of practice.
- Have observers discuss their beliefs and implicit theories of practice.
- Ask them how their beliefs and implicit theories might influence how they record and evaluate evidence.
- Warn observers to be aware of potential biases and to focus on and rate using the specific definitions and explanations of the rating scale.

4. Explain common rater errors to be aware of and avoid.

- Similarity – rating influenced by how similar the observed classroom or school is to yours, how similar the practice observed is to yours, or how similar the person being observed is to you.
- Leniency – rating higher than deserved to give the person the "benefit of doubt."
- Halo – rating on one dimension determined by rating on another.
- Central tendency – rating everyone in the middle; often due to "anchoring" on the middle level by assuming that everyone is average (or proficient) unless there is a lot of evidence he/she is not.
- Consistency/confirmation – looking for evidence for pre-judgment or a judgment based on one's initial impression.
- Context effects – performance of peer group influences ratings. Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings

5. Describe the process for decision-making.

- Emphasize separating the observation (or other evidence collection) from the judgment about the level of practice (which is based on comparing the evidence to the rubric or rating scale).
- When taking notes, record what was observed in behavioral terms.
- Do not rate while observing.
- Review notes after finishing observation; highlight evidence that is relevant to each dimension.
- Compare performance observed to the rubric or rating scale, not to other performers.
- Respect the rubric over your gut feeling. (Don't rely on 'I know good teaching when I see it.')
- Evaluate based only on the evidence collected: if no evidence, make no inference.
- Where evidence is mixed on whether observed performance meets the requirements for rubric level, base decisions on the predominance of evidence. If a substantial majority of the evidence supports rating at a specific level, choose that level rather than the level below.
- Avoid anchoring – assuming the performance is satisfactory or proficient unless there is evidence to the contrary.
- Rate performance on each dimension or standard separately.
- Try not to compensate for a near miss on one dimension with a generous rating on another.

6. Have observers practice observing and recording evidence; discuss and provide feedback to observers.

7. Have observers practice connecting evidence recorded from the observation to performance dimensions.

- Discuss questions about what performance standards or dimensions cover.
- Review rubrics: what am I looking for?
- Review notes/artifacts and identify evidence related to rubric dimensions.

8. Have observers practice interpreting the rubrics.

- Identify the specific rubric language that differentiates between different performance levels.
- Discuss questions observers may have about the interpretation of rubric language.
- Review rating techniques and conventions (e.g., how a word like 'consistently' is to be interpreted).
- Practice rating using videos, written scenarios, or live observations.
- Have observers share ratings, discuss reasons for ratings; trainer then provides feedback to observers on how well they are doing.
- Repeat for all rubric dimensions or standards.

9. Rater training may be followed by a 'certification exercise' in which evaluators must match the ratings of videos, observations, or artifacts done by expert jury in order to be allowed to do assessment in the field. Usually some threshold is set, such as 75% absolute agreement with the experts. Trainees who fail are retrained.

   This includes developing a shared mental model of good performance first among the observers and then among the educators being observed. (*ibid.*, at pp. 25 – 27)

9.4.2 Ontario, Canada

Ontario is a province of Canada, which has responsibility for education policy under that country's federal arrangements. The Education Quality and Accountability Office (EQAO) is an independent provincial agency funded by the Government of Ontario. EQAO's mandate is to conduct province-wide tests at key points in every student's primary, junior and secondary education and report the results to educators, parents and the public (EQAO, 2012).

A technical manual for Assessments of Reading, Writing and Mathematics, in the Primary Division (Grades 1–3) and Junior Division (Grades 4–6); Grade 9 Assessment of Mathematics and the Ontario Secondary School Literacy Test in 2010 - 11 is available online (*ibid.*). This manual has appendices describing various aspects of task scoring. The range-finding process for open-response items defines the range of acceptable performances for each scoring point in each rubric that is used to train scorers. Range finding precedes field testing (*ibid.*, at p. 12). Detailed procedures are in place for field test scoring, including training raters (*ibid.*, at pp. 14 – 15).

In addition, there are many controls on operational scoring. The 2011 technical manual describes the room in which scorers work in detail (*ibid.*, at p. 16). It also has the following description of training procedures:

9.4.2.1 Training for scoring open-response operational items

> The purpose of training is to develop a clear and common understanding of the scoring materials so that each scoring leader, scoring supervisor and scorer applies the scoring materials in the same way, resulting in valid (accurate) and reliable (consistent) student scores.

9.4.2.2 Training of scoring leaders and scoring supervisors for scoring open-response operational items

Scoring leaders must have subject expertise and be, first and foremost, effective teachers of adults. They must encourage scorers to abandon preconceived notions about scoring procedures and align their thinking and judgment to the procedures and scoring materials for the items being scored. The responsibilities of scoring leaders include

- training all scoring supervisors and scorers in the applicable room;
- overseeing the scoring of items;
- ensuring that scoring materials are applied consistently and
- resolving issues that arise during scoring.

Scoring leaders are also responsible for reviewing and analyzing daily data reports to ensure that a high quality of scoring occurs in their scoring room.

Scoring supervisors are selected from a pool of experienced and proficient EQAO scorers. Scoring supervisors assist scoring leaders and ensure that their assigned scorers are qualified and are scoring accurately. Scoring supervisors may also be asked to retrain individual scorers when necessary.

The training for scoring leaders and scoring supervisors is conducted before scoring begins. EQAO education officers train scoring leaders and oversee the training of scoring supervisors. Supervisor training is substantially similar to the training and qualifying for scorers. The only difference is that supervisors receive additional training regarding scoring materials, room management problems and issues that may arise during scoring. For Grade 9 scoring, an EQAO education officer trains the scoring leaders and supervisors assigned to one room at the same time.

Following training and prior to scoring, scoring leaders and scoring supervisors must pass a qualifying test that involves scoring 14–20 student responses for the items to be scored in their room. The items included in the qualifying test are selected during the range-finding process.

Scoring leaders and supervisors must attain at least an 80% exact and a 100% exact-plus adjacent match with the expertly assigned scores. Scoring leaders or supervisors who fail the qualifying test may not continue in the role of leader or supervisor.

9.4.2.3 Training of scorers for scoring open-response operational items

The purpose of training for open-response operational items is to ensure that all scorers become experts in scoring specific items or subsets of items. All operational items require a complete set of scoring materials: generic or item-specific rubrics, anchors (real student responses illustrating work at each code in the rubric) and their annotations, training papers, a qualifying test, validity papers (primary, junior, OSSLT) or validity booklets (Grade 9) and items for the daily calibration activity.

To obtain high levels of validity (accuracy) and reliability (consistency) during scoring, EQAO adheres to stringent criteria for selecting, training and qualifying scorers. Various other quality control procedures, as outlined below, are used during the scoring process to identify scorers who need to be retrained or dismissed from scoring.

All the scorers in one room are trained to score using the same scoring materials. These scoring materials are approved by EQAO and cannot be altered. During training, scorers are told they may have to adjust their thinking about scoring student performance in a classroom setting in order to accept EQAO's standards and practices for its assessments.

Training for scorers on the limited number of open-response items scored in a room takes approximately half a day and includes

- general instructions about the security, confidentiality and suitability of the scoring materials;
- instructions on entering scores into the Personal Digital Assistant (PDA) used to collect scoring data. For instance,
    - prior to entering scores, scorers scan the unique student booklet barcodes into the PDA (which has a built-in barcode scanner) in order to link student names to their corresponding scores and
    - scorers enter their scores for student responses into the PDA, then synchronize the PDA in a cradle connected to a laptop, which uploads the data to a server;
- a thorough review and discussion of the scoring materials for each item to be scored (the item, generic or item-specific rubrics, anchors and their annotations):
    - emphasis is placed on the scorer's understanding of how the responses differ in incremental quality and how each response reflects the description of its code on the rubric and
    - the anchors consist of responses that are typical of each achievement level (rather than unusual or uncommon) and solid (rather than controversial or 'borderline') and
- the scoring of a series of validity papers or validity booklets (Grade 9), consisting of selected expertly scored student responses:
    - validity papers or validity booklets (Grade 9) typically contain responses that are solid examples of student work for a given response code. Scorers will first score the responses and then synchronize the PDA and
    - scorers will then discuss the attributes and results of each correct response with their scoring leader and supervisor. They will internalize the rubric during this process and adjust their individual scoring to conform to it.

Scorers are also trained to

- read responses in their entirety prior to making any scoring decisions;
- view responses as a whole rather than focusing on particular details such as spelling;
- remain objective and fair and view the whole response through the filter of the rubric and
- score all responses in the same way, to avoid adjusting their scoring to take into account a characteristic they assume about a student (e.g., special education needs, being an English language learner).

Following training and prior to scoring, scorers must pass a qualifying test consisting of 14–20 student responses to all the items to be scored in a room. These items are selected during the range-finding process as examples of solid score points for rubrics. Scorers must attain at least a 70% exact match with the expertly assigned score. This ensures that scorers have understood and can apply the information they received during training. Scorers who fail the qualifying test the first time may undergo further training and write the test a second time. Scorers who fail to pass the qualifying test a second time are dismissed. (*ibid.* at pp. 16 – 18)

In addition to these detailed training procedures, the EQAO manual describes how marking is monitored on an ongoing basis, referring to a concept of 'daily and cumulative validity', which is monitored using a range of statistical indicators (*ibid.,* at p. 20).

9.4.3 Texas, USA

The Texas English Language Proficiency Assessment System (TELPAS) is designed to assess the progress that limited English proficient (LEP) students make in learning the English language. There is extensive information about TELPAS rater training procedures online, and the following passage summarises some of the ground rules for rater training that are expressed online.

TELPAS training can be summarised in this figure:

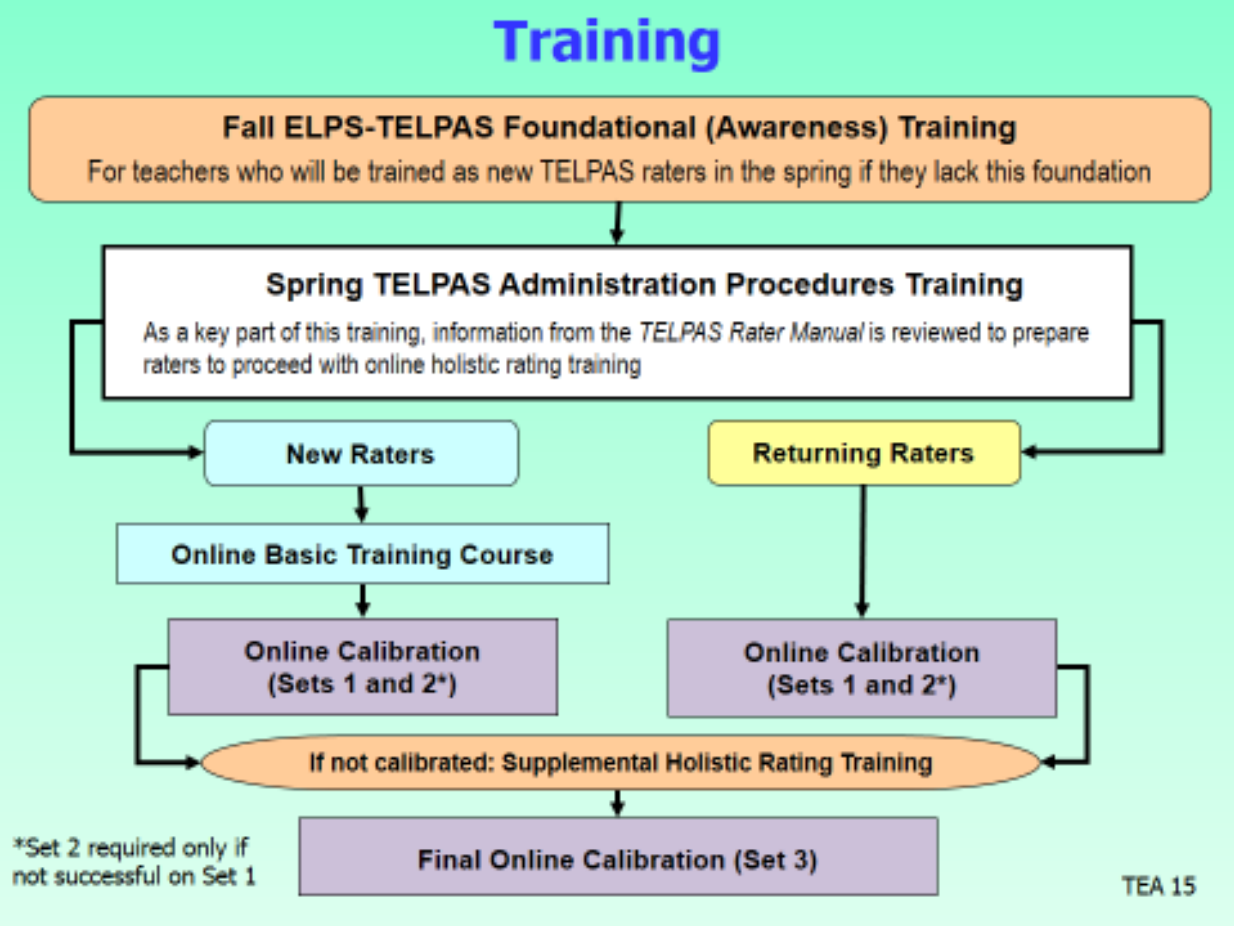

Figure 1: Summary of TELPAS training procedures (TEASAD, 2013a)

The purposes of online basic training and online calibration can be descriptive, respectively, as follows:

- **Online basic training course**

  This course is for new raters. It provides instruction on using the rubrics and gives raters practice rating students in each language domain. There are separate courses for K–1 and 2–12.

- **Online calibration**

  This is for all raters. Raters use the rubrics to rate students in each language domain. Raters have three opportunities to calibrate on assigned grade cluster. (*ibid.*)

The content of these training events is as follows:

> **Online Basic Training Course (*Required for New Raters*)**
>
> There are two basic training courses, one for raters of K–1 students and one for raters of students in grade 2 or higher (2–12). The K–1 course covers the four language domains of listening, speaking, reading, and writing. The 2–12 course covers listening, speaking, and writing. After learning the basics of the holistic rating process, participants practice rating students as part of the course. New raters must complete this course before beginning online calibration activities. Approximate completion time: 4–5 hours.
>
> **Online Calibration (*Required for New and Returning Raters*)**
>
> The online calibration activities consist of three sets of students to be rated. Each language domain is represented in each set. For K–1, each set includes all four language domains—listening, speaking, reading, and writing. For 2–12, each set includes listening, speaking, and writing. Raters complete only as many sets as it takes to calibrate. Approximate completion time per set: 1 hour. (TEASAD, 2013b, p. 14)

Raters can take the online calibration either in their own homes, or schools. The need for annual re-calibration is justified to rates as follows:

> Standardized testing programs include processes to ensure that all individuals assessing students interpret the scoring rubrics the same way. Scorers of written compositions for the STAAR program complete calibration activities.
>
> Yearly calibration is a necessary aspect of administering a holistically scored assessment. When holistic assessment processes are used, even the most experienced scorers need to make sure they are calibrated to score accurately.
>
> Over time, calibration activities serve to give raters more examples that help expand their knowledge and help them rate students who are near the border between two proficiency levels or who exhibit less typical language characteristics. (TEASAD, 2013a)

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

First 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