No news is good news?
Talking to the public about the reliability of assessment

Andrew Boyle, Dennis Opposs and Annette Kinsella

Presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference in Brisbane, Australia, 13–18 September, 2009

August 2009
Ofqual/09/4362

Abstract

Quantifying the reliability (or unreliability) of assessment results is a central part of the work of assessment professionals. Although much remains to be done, methods for calculating and reporting reliability indices have been widely discussed.

Communicating with the public about unreliability in test scores has not been addressed to the same extent. In its initial public communications on this, Ofqual, the regulator of examinations and qualifications in England, has found that unreliability is a difficult topic for two reasons. Firstly, the concept of reliability is complex and hard to explain succinctly. Secondly, unreliability seems like an intrinsically bad news story.

This paper will report on two sources of evidence. Firstly, literature describing the media environment that surrounds examination results in England will be summarised. Such literature can give a history of assessment organisations' attempts at communicating with the public, and make suggestions for how such bodies might communicate better. The second source of evidence is a series of workshops conducted to Ofqual's specification by UK social research organisation Ipsos MORI. That work has given Ofqual an initial feel for the tolerance that different sectors of the public have for different sources of measurement inaccuracy in examination results.

The paper will conclude by suggesting ways to improve each of the issues with unreliability as a media story; the problem of complexity will be addressed by allowing people to interact with the message via multiple media, using varied analogies and so on. In terms of the negativity of the story, the response will not be to try to make this into a good news story. Rather, the aspiration is to communicate the message that many assessment results contain an element of unreliability to the public in a manner that allows people to become more sophisticated users of those results.

Introduction

Reliability

Reliability is a fundamental property of measurement provided by examinations, qualifications and tests. Indeed, reliability is a key indicator of the quality of information that such instruments provide. Reliability can be defined as:

> … the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability means that broadly the same outcomes would arise. (Newton, 2009)

Reputable sources suggest there is a duty to communicate about the reliability of results to the public. (See, for example, the *Standards for Educational and Psychological Testing* – AERA et al, 1999 – standard 2.1, p. 31.)

Reliability is not the principal, nor the most frequently debated, indicator of the quality of assessment results. Current conceptualisations describe validity as the fundamental indicator of the quality of the evidence to support an assessment's intended use. Reliability is seen as a facet of validity or a source of evidence that can contribute to a validity argument (AERA et al, 1999, p. 17). In England, there have been several public controversies around examinations and national tests in recent years (Richardson, 2007; Hood et al, 2007), but these have tended to focus on maintaining standards over time and/or comparability between assessment organisations or subjects (Newton et al, 2007), rather than reliability.

The assessment scene in England

There are several major assessment occasions in the English education system. There are full-cohort national tests – previously these were in English, mathematics and science at ages seven, 11 and 14, now they are taken by seven and 11-year-olds; at the end of compulsory schooling students sit qualifications, including the General Certificate of Secondary Education (GCSE); in post-compulsory schooling there is the General Certificate in Education, Advanced level (GCE A level). There is also a wide range of vocational and occupational qualifications.

Purpose of this paper

The purpose of this paper is to consider issues that affect how organisations with responsibility for assessment can best communicate with the public. This specifically refers to the major technical issue of reliability, describing the approach that is being taken by the examinations and qualifications regulator, Ofqual, in a large programme of work in England. The paper will reflect upon what the work on communicating about reliability tells assessment organisations about how to communicate about technical issues in general.

Structure of this paper

The final part of this introduction contains a selective review of literature on education and the media. After that, findings from empirical research conducted to understand some sections of the public's opinions about reliability are reported. The paper concludes with a discussion section, which reflects on lessons learned about communicating with the public about this complex area.

Literature on education and the media

In previous decades, education was a less prominent political issue. For example, Margaret Thatcher subsequently became the UK Prime Minister, but her biographer states that she knew (in 1970, when appointed to the post) that Secretary of State for Education wasn't a 'mainline political job' (Young, 1989, p. 67). Since then, education has increased substantially in political importance (Thrupp & Tomlinson, 2005) and educational standards in particular have become a matter of partisan debate. With this politicisation, examination standards have greatly increased as a subject of media stories (Warmington & Murphy, 2004, p. 289).

It is instructive to compare and contrast the business of assessment provision and broadcast media; one version emphasises the difference between the two professions and businesses (to the benefit of assessment provision) and the other emphasises the similarities between the two.

In the first comparison the differences between providing examinations, qualifications and tests on the one hand and writing and publishing journalistic stories is emphasised. According to this analysis, the provision of valid and reliable assessment outcomes is an ethical activity; organisations, and the researchers who work for them, are engaged in scientific investigation of a complex reality. By doing so, they seek veracity, fairness and consistency. A central concept within reliability theory is that of 'true score' (Feldt & Brennan, 1989). In detrimental contrast, journalism seeks simplicity of content and straightforward narrative structure (beginning, middle and end) (Levin, 2004). Media stories emphasise the short term and 'thrive on wrong doing' (*ibid*, p. 279); they seek definitive conclusions and do not have the time and space for nuanced debate. Assessment professionals might speak dispassionately about 'error' in examination results as an inevitable feature of psychological measurement, but journalism would seek to portray such error as the fault of an agent, and if such an agent could be constructed as a remote, unconcerned bureaucrat then so much the better.

The second analysis emphasises similarities between journalism and the provision of high-stakes assessment. Journalism is a profession whose ethics, aesthetics and practices (eg reliance on and protection of the confidentiality of sources) reflect a commitment to truth telling that overrides commercial considerations (McNair, 1998).

A central contribution of journalism to the greater good of society is to reveal truths and thus hold powerful executive forces to account.

Media and assessment organisations have both undergone substantial and similar changes in recent years. Media organisations have experienced a flight from the local to the national and international level. Commercial realities have led to many regional newspapers being closed (Luft, 2009), and local TV and radio news services have come under pressure (Ofcom, 2008). The previously regional exam boards in England were consolidated into three national unitary boards (providing both 'general' and 'vocational' qualifications) in a long-running process culminating in the 1990s (Tattersall, 2007). This was mainly done for technical rather than commercial reasons (concerns about a lack of comparability between a proliferation of organisations), but England's providers of general examinations do compete against each other in an admittedly imperfect market (Europe Economics, 2008, p. 15)[1].

Media and assessment organisations are both influenced by the rise of technology and globalisation. The presence of online and international news providers has already affected UK media organisations profoundly, whereas it has as yet only done so superficially for assessment organisations and similar bodies. However, factors such as on-demand testing and the provision of qualifications by offshore organisations are likely to impact on assessment organisations sooner rather than later (Ridgway, McCusker & Pead, 2004).

Thus, in comparing assessment and news provision one may make two analyses. The first emphasises the differences between the two to the advantage of the former, whereas the second emphasises features that are held in common. At this stage, one cannot favour either analysis, except to say that simplistic contrasts should be avoided. Indeed, thinkers such as Foucault (1972, p. 12) suggest that false juxtapositions can be used to denigrate 'the Other' as a way of indirectly and surreptitiously constructing actors or disciplines as virtuous, logical and above criticism. Although he is not coming from a Foucauldian perspective, it is clear that Baker (2000) answers his own question 'Does education get the media it deserves?' with a resounding 'yes'.

Newton has written about assessment organisations' obligations to communicate in public about measurement issues (Newton, 2005a and Newton, 2005b). He defines terms carefully. He refers to 'measurement inaccuracy' as 'the variety of ways in which *any* set of assessment results will *always*[2] depart from the mythical ideal of perfect accuracy for all students' (Newton, 2005a, p. 420). 'Measurement inaccuracy' is a broadly conceived notion and includes reliability, validity and comparability

[1] The qualifications market is 'imperfect' in that those in charge of purchase decisions typically do not use price as the key determiner for their purchase.
[2] Emphasis in original.

deficits (*ibid*). It can be contrasted with ‘human error’, which includes ‘head-slappingly obvious mistakes’ (*ibid*). Human errors are, for practical purposes, inevitable in large-scale testing programmes, but they are not inherent as a matter of principle – in contrast to measurement inaccuracy (*ibid*, p. 421). Newton (2005b, p. 458) uses the overarching term 'assessment error' to include both 'measurement inaccuracy' and 'human error'.

Newton (2005a) surveys literature concerning trust and distrust, both generally concerning governments’ actions and specifically with respect to examinations provision. He cites thinkers who have alleged that there is a ‘crisis’ of trust, although he also notes that people sometimes appear to trust institutions in fact (by their actions) even when they state that they do not trust them (2005a, p. 422). Newton debates whether trust and transparency are mutually exclusive; citing the possibility that the public might need to be given less information, lest people be perturbed by finding out the truth about measurement inaccuracy (*ibid*, p. 426). Newton also looks at the relationship between measurement inaccuracy and culpability, and suggests that there are two potential logical fallacies; firstly, that measurement inaccuracy (necessarily) implies culpability (for instance in an assessment agency), and secondly, that lack of culpability (for instance when measurement inaccuracy is inherent) implies accuracy (Newton, 2005a, p. 430). He suggests that the first fallacy is widely held amongst the general public, and the second amongst professionals in assessment organisations.

Newton concludes that assessment organisations should be transparent and communicate with the public about measurement inaccuracy. He does so both on ethical and practical grounds. He states that assessment organisations have a duty to provide test users with sufficient information to allow them to make valid inferences based on test results (*ibid*, p. 431), but he also points out that – since measurement inaccuracy exists in fact – it would be rather ‘tactically naïve’ of agencies to deny its existence, because to do so would be to concede the point on culpability. That is, doing so would not challenge the suggestion that agencies were to blame for all measurement inaccuracy (*ibid*, p. 433). He ends, optimistically, by proposing that it is possible to have a proactive programme of communication and public understanding of measurement inaccuracy that would be beneficial both from the perspective of improving the ethical conduct of assessment organisations and making their job easier by acquainting the public with the truth about inaccuracy.

Method

Ofqual is the regulator of examinations, qualifications and tests in England. It is being set up by legislation to be independent of the government education ministry, and instead to be answerable directly to the UK Parliament (Ofqual, 2008). One of Ofqual’s substantial early programmes is addressing reliability (Opposs, 2009). That programme has been divided into three strands:

- Strand 1: generating evidence on reliability
- Strand 2: interpreting and communicating evidence on reliability
- Strand 3: exploring public understanding of reliability and developing Ofqual policy on reliability.

A range of projects have been commissioned in strands 1 and 2, and reports of their findings will be posted at www.ofqual.gov.uk/1999.aspx in due course. The current paper reports on two projects carried out under strand 3.

The first project was an attempt to gain initial insights into the public's understanding of, and attitudes to, reliability in assessment results. The second, which was undertaken in response to the experience of the first, attempted to develop a succinct but meaningful narrative that Ofqual could use when communicating with the public about reliability.

The first exercise to investigate the opinions of several sections of the public in relation to reliability or unreliability in examination results was conducted by the social research company Ipsos MORI to Ofqual's commission (Ipsos MORI, 2009). This research sought the opinions of the following groups: teachers (of secondary education students), students (aged 16–19 and in secondary education), parents (with children in secondary education, years 7–13), members of the general public (either with no children, young children under 10, or older children aged 20+) and employers (*ibid*, p. 6).

The research was conducted using a workshop methodology. This approach involved session facilitators providing more substantive input to participants (for example by setting up analogies to illustrate the issue of reliability and unreliability in test scores) than would be the case in some other research methods (such as focus groups) (see discussion guides in Ipsos MORI, 2009, pp. 39ff). This approach was taken because it was felt (prior to the field research) that participants might well not have developed opinions about the issue under discussion. Therefore information on the topic was provided to participants to help them to develop views on reliability. It was understood that by providing substantial input to participants, the research ran the risk of biasing the participants' views. However, it was felt that this risk was less serious than the risk that participants might not have any view about inaccuracy in exam scores or grades.

The second piece of work was developed after Ofqual staff reflected on the experience of running the first opinion-gathering exercise. It consisted of a session with the communications messaging consultant Blue Rubicon, which was used to produce a narrative for Ofqual staff to use when speaking about reliability and unreliability. The spur for this work came from the observation that it was not easy to express ideas around reliability in terms that were informative yet consistent, concise

and comprehensible. This was particularly an issue when different members of staff (for instance communicators, researchers and policy makers) needed to speak about reliability, or when third parties (for instance consultants or contractors) needed to do so.

Narratives of this type are often used as part of campaigns – for instance by companies promoting a product or service, or by political parties or other campaigning organisations. However, in this instance no campaign was being undertaken except – perhaps stretching the term – a public information campaign; trying to help the public to become more informed about reliability and unreliability.

Participants

The research into the opinions of different groups about reliability and unreliability consisted of two workshops held in London and Birmingham in January 2009. There were 36 participants in each workshop, split up as follows:

- eight teachers
- six students
- six parents
- six members of the general public
- six employers
- four examiners.

(Ipsos MORI, 2009, p. 6)

The messaging workshop was conducted at Ofqual's London office and facilitated by Blue Rubicon staff, with a range of Ofqual policy makers, communicators and researchers attending.

Instrumentation

The workshops on public perceptions were conducted according to detailed yet flexible discussion guides. Ipsos MORI researchers and Ofqual staff discussed their development in detail and the guides are reproduced at (Ipsos MORI, 2009, pp. 39–45).

The workshop started with the session facilitator invoking the analogy of medical misdiagnosis as follows:

> What could go wrong when a medical diagnosis is made? **PROMPT** Think about the patient? And the doctor? What about the instruments the doctor uses?
> Pull out issues that relate to the different contributions of:
> THE PATIENT:
> What if they describe a lot of co-incidental symptoms unrelated to the real problem?

> What if they do not describe their symptoms accurately enough?
> What if their symptoms are not present on the day of their appointment?
> What if their symptoms are not severe enough to make the doctor aware of how unwell they really are?
> THE DOCTOR:
> What if they are pre-occupied with the previous patient, or under pressure to finish their clinic and fail to diagnose the problem?
> What if they are not up-to-date in their training on this particular condition?
> What if they simply misinterpret the symptoms in front of them and make an incorrect judgement?
> THE INSTRUMENTS
> What if the sample of blood or cells taken for diagnosis becomes corrupted?
> What if the thermometer is broken or the sensitivity of the heart monitor is set incorrectly, even if the doctor is using the instruments correctly?
> What if the instruments themselves are poorly designed, or broken?
> In each case;
> Who is responsible for the problem?
> What are the implications of the problems?
>
> (Ipsos MORI, 2009, p. 40)

As stated above, it was understood that giving such substantial input to research participants whose opinions and attitudes one was trying to discover ran the risk of biasing them. However, the belief was that participants would probably not have developed views on reliability in test scores and so it was felt important to give them contextualisation of this sort.

On the advice of Ofqual staff, the Ipsos MORI session facilitators used the term 'error' throughout, rather than 'measurement inaccuracy', 'reliability' or any other term.

Findings

Qualitative research into stakeholders' opinions

The findings suggested a demarcation in the minds of the public between inevitable errors in the assessment process and preventable errors. The research participants appeared to accept that a certain amount of error was inevitable in a large examinations system, but they could be intolerant of 'preventable errors' (Ipsos MORI, 2009, p. 15). However, these findings need to be interpreted carefully, especially if the wish is to confirm or refute Newton's earlier thinking on 'measurement inaccuracy' and 'human error'. It is far from clear that those concepts were the strongest explanators of the variations in respondents' opinions. Rather, it is at least arguable that differences in opinions can be understood more clearly by referring to the group to which the opinion-holder belongs (teacher, student, parent, employer, examiner), the perceived agent of the error (examiner, exam board, student) and the consequence of the error. Sometimes participants appeared to be making a distinction between inherent and preventable error, but other times not. Also, culpability and assessment error appeared to be entwined issues.

Some teacher and employer participants in the research stated their differential attitude to error depending upon whether the error changed a student's grade or

mark (Ipsos MORI, 2009, p. 16). They considered grade-related error to be more consequential than mark-related (*ibid*). Participants' views about error could vary by group, and by the perceived cause of the error. For example, students and teachers could be intolerant of typos in papers (Ipsos MORI, 2009, p. 23–4), while examiners could be more sanguine – taking the view that what was important was that any mistakes that did occur were rectified (*ibid*).

The findings on 'examiner-related error' show how the various strands are intertwined. For example, there is evidence that students are aware that some inconsistency between human markers is inherent in subjects such as English (*ibid*, p. 21). However, there are also statements that such inherent error should be minimised or even eliminated. Some participants suggest practical measures such as the double marking of papers or making markers do their work in marking centres rather than at home (*ibid*, p. 22).

There is considerable thoughtful discussion on 'test-related error' (Ipsos MORI, 2009, p. 24). Students and the general public are able to debate whether and how examinations can and should sample from curricula (*ibid*). They even go on to evaluate the merits of the different assessment methods: terminal examinations compared to modular assessment and coursework. This is a sophisticated debate about the validity of qualifications systems, and not a reduction to 'head-slappingly obvious error'.

The final finding from the Ipsos MORI research to be reported here concerns the word 'error'. The researchers reported that this term had some negative impacts when used with the public (Ipsos MORI, 2009, pp. 37–8). In particular, the common meaning of that word, in contrast to its technical meaning, reinforced any latent disposition to treat unreliability as necessarily imputing culpability. Further, the word grammar of 'error' tends to cause the issue of inherency, agency and culpability to be further muddied. For example, to speak of 'an error' (a count noun) seems to imply a single event, for which some person or thing must be responsible/culpable. In contrast, the slightly less common, more 'scientific' use of 'error' as a non-count or mass noun lessens the necessary connection with culpability. This degree of syntactical subtlety and potential for ambiguity suggests that this is not an ideal word to use centrally in an important public communication campaign.

Messaging workshop

The Blue Rubicon/Ofqual discussion session produced a two-page narrative document. The narrative consists of the following elements:

- fundamentals – why the programme exists
- strategy – Ofqual’s response to the fundamentals

- proof – of delivery of the strategy with evidence of progress to date
- vision – for the reliability programme itself and related areas.

The narrative describes the programme and its strands. It mentions projects that have been commissioned and claims that:

> This programme is a major undertaking and a significant contribution to a better understanding of the qualifications system more generally. (Blue Rubicon, 2009)

The document also provides an agreed set of terms with which to refer to key concepts in the programme. In particular, it settles on the term 'variation' (in scores, assessment procedures, etc) to describe the thing the reliability programme is talking about.

It was necessary to choose an alternative term to 'error' as this was too closely associated with culpability, and because it had an unhelpfully subtle word grammar. 'Variation' was felt at the session to be the best candidate to use, ahead of alternatives such as variance, uncertainty, discrepancy, inconsistency or clash. It is possible that some of these terms could also be used, while others would not be useful. 'Variance' would probably not be a good candidate, since it is confusingly associated with a statistical concept (which has a certain, but not complete, relationship to reliability). 'Clash' is probably not close enough in meaning to unreliability and also has the potential to provide incendiary headlines. However, members of the reliability programme could try out some or all of the other terms when speaking in public about reliability.

Discussion

This final section of the paper reflects on the findings that are available on the ‘communicating with the public strand’ of the Ofqual reliability programme, which is currently about halfway through its full duration. To do this, the paper sets up a series of diptychs or juxtapositions to illustrate some of the choices and dilemmas that confront the programme.

The first juxtaposition is between measurement inaccuracy, which is said to be inherent in principle and therefore excusable on the part of assessment organisations. This is contrasted with (‘head-slappingly obvious’) human error, which is less excusable, but more obvious to the non-expert eye. There is evidence from the workshop discussions reported in this paper that members of the public were able to have sophisticated discussions of measurement inaccuracy, and to be able to propose and evaluate steps that would minimise it (for instance, double marking or stricter controls on markers). This is a positive finding for a ‘democratic’ view of communicating with the public; people are entirely capable of taking on board the most complex forms of measurement inaccuracy. It also puts a logical limit on the inherency of measurement inaccuracy. It is not possible to entirely eliminate all forms

of measurement inaccuracy, but most of them can be substantially reduced if there is a public will to do so. This work may tentatively suggest that the costs and benefits of reducing 'inherent' measurement inaccuracy can be discussed in an informed, legitimate and democratic manner. If public will to mitigate causes of measurement inaccuracy could be established with an informed understanding of concomitant costs, this would suggest that technical features of assessment could be improved alongside a corresponding increase in public legitimacy of the examinations or qualifications concerned.

A second contrast set up in the introduction to this paper was between the supposedly scientific business of assessment provision and the more mundane activity of journalism and the provision of news. By the end of this research exercise using this contrast to provide a favourable description of assessment provision seems increasingly difficult to sustain. There are many scientific disciplines (climate science, stem cell research, the economics of downturns) that must argue their case in the popular media. It seems entirely reasonable that agencies involved in the provision of examinations, tests and qualifications should argue their corner through the media that are available. Indeed, it is likely that other communications techniques such as multimedia presentations, website discussions and face-to-face discussions could be used profitably to put across messages around reliability.

However, in saying this, one must also reflect on challenges. It seems legitimate, if novel, for assessment researchers to use journalistic or communications techniques. But as they do so they must remain critical of the activity they are engaged in. In particular, care should be taken to separate out the uses of journalistic techniques for campaigning (in favour of products or ideas) from those of providing public information. It seems legitimate to use these techniques to provide clear information to the public, but less so to engage in campaigning.

Somewhat similar is the contrast and dilemmas involving educating the public about reliability and seeking their opinions. It may well be that educating the public about rather esoteric aspects of reliability theory is an illegitimately top-down or patrician conceptualisation of the activity that regulators and assessment organisations should be engaged in. The Ipsos MORI research does – in places – support the view that the public can formulate and debate sophisticated validity arguments. Moreover, Newton's concept of measurement inaccuracy covered reliability, validity and comparability. This may mean that, in seeking the public's views about inaccuracy, researchers will have to address a dilution from the purist notion of reliability. This seems preferable to seeking to 'educate' respondents in arcana of reliability theory.

A related point concerns the amount of support or input that may be given at the start of an opinion-gathering exercise without unreasonably biasing respondents' opinions. The workshop format used to date does require substantial information to be given to participants, beyond what would be given in approaches such as focus groups. This

is not unique to the current research context, however. Somewhat similar issues exist in international questionnaire research, where respondents from around the world may struggle to pick up the implicit assumptions made by a researcher from a different culture. In such contexts, it can be legitimate to provide substantial information to respondents – for instance by giving the model answer that the researcher's country would give to a particular question (Boyle, 2008).

The need to support respondents in clearly understanding the issue that research questions address may be more substantial when confirmation of opinions is sought from a representative sample of respondents. Instruments to get such large samples of opinion such as written questionnaires need to be comprehensible and robust, without requiring respondents to consume large quantities of complicated information before they are able to give their views. It is suggested that this dilemma might be resolved by using relatively broadly conceived questions about measurement inaccuracy and variation, the use of some concise, well-written exemplification and sophisticated post hoc analysis to help understand how questionnaire items have performed. There is some irony in the reflection that such analysis would be highly likely to include an evaluation of the reliability of the data set generated by the questionnaire.

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing.* (Washington, DC: American Educational Research Association).

Baker, M. (2000) *Does education get the media it deserves? An inaugural lecture*. (London: Institute of Education).

Blue Rubicon (2009) Ofqual reliability programme messaging, forthcoming, see: http://www.ofqual.gov.uk/reliabilityprogramme.

Boyle, A. (2008) *The regulation of examinations and qualifications: an international study*. Available online at: http://www.ofqual.gov.uk/international-regulation.

Europe Economics (2008) Scoping exercise for a study of the efficiency of the qualifications system: a report for QCA. Available online at: http://www.ofqual.gov.uk/files/scoping_study_october_29_2007.pdf.

Feldt, L.S. & Brennan, R.L. (1989). Reliability in R. L. Linn (Ed.) *Educational Measurement*. (Washington, DC: American Council on Education/Macmillan). 3rd edition.

Foucault, M. (1972) *The archaeology of knowledge*. (London: Routledge).

Hood, C., Jennings, W. and Hogwood, B. with Beeston, C. (2007) *Fighting fires in testing times: exploring a staged response hypothesis for blame management in two exam fiasco cases*. London School of Economics/ESRC Research Centre discussion paper no: 42. Available online at: http://www.lse.ac.uk/collections/CARR/pdf/DPs/Disspaper42.pdf.

Ipsos MORI (2009) *Public perceptions of reliability in examinations*. Available online at: http://www.ofqual.gov.uk/files/2009-05-14_public_perceptions_of_reliability.pdf.

Levin, B. (2004) Media-government relations in education, *Journal of Education Policy*, 19(3), pp. 271 – 283.

Luft, O. (2009) *Trinity Mirror to close nine Midlands newspapers*. Available online at: http://www.guardian.co.uk/media/2009/jul/01/trinity-mirror-to-close-midlands-papers.

McNair, B. (1998) *The Sociology of journalism.* (New York: Arnold).

Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (Eds.) (2007) *Techniques for monitoring the comparability of examination standards*. (London: QCA).

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

Newton, P.E. (2005b) Threats to professional understanding of assessment error, *Journal of Education Policy*, 20(4), pp. 457 – 483.

Newton, P.E. (2009) The reliability of results from national curriculum testing in England. *Educational Research*, 51(2), June 2009, p. 181 – 212.

Ofcom (2008) *Local TV and public service broadcasting*. Available online at: http://www.ofcom.org.uk/media/features/salford.

Ofqual (2008) *Introducing the new regulator of qualifications, exams and tests*. Available online at: http://www.ofqual.gov.uk/files/Ofqual_LaunchBrochure.pdf.

Opposs, D. (2009) *Ofqual's reliability of results programme*. Paper presented at the Chartered Institute of Educational Assessors' third National Assessment Conference, London, Wednesday 6 May 2009. Available online at: http://www.ciea.org.uk/upload/conference_2009/presentations/seminar%203.ppt.

Richardson, W. (2007) Public policy failure and fiasco in education: perspectives on the British examinations crises of 2000-2002 and other episodes since 1975. *Oxford Review of Education*, 33(2), pp. 143-160.

Ridgway, J., McCusker, S. and Pead, D. (2004) *Literature review of e-assessment*. Available online at: http://www.nestafuturelab.org/research/reviews/10_01.htm.

Tattersall, K. (2007) A brief history of policies, practices and issues relating to comparability in Newton, P., Baird, J-A., Goldstein, H., Patrick, H. & Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards.* (London: QCA).

Thrupp, M. & Tomlinson, S. (2005) Introduction: education policy, social justice and 'complex hope'. *British Educational Research Journal*, 31(5), pp. 549–556.

Warmington, P. & Murphy, R. (2004) Could do better? Media depictions of UK educational assessment results. *Journal of Education Policy*, 19(3), pp. 285 – 299.

Young, H. (1989) *One of Us*. (London: Pan/Macmillan).

All web links were accessed on July 13 2009.