Cite as: Sinkinson, A. and Jones, K. (2000), The Validity and Reliability of OFSTED Judgements of the Quality of Secondary Mathematics Initial Teacher Education Courses. Paper presented at the Symposium on 'Critical Issues in Mathematics Initial Teacher Education' at the *British Educational Research Association Annual Conference* 2000 (BERA2000), The University of Wales, Cardiff, 7 – 9 Sept 2000.

The Validity and Reliability of Ofsted Judgements of the Quality of Secondary Mathematics Initial Teacher Education Courses

Anne Sinkinson, Homerton College, Cambridge,
and
Keith Jones, University of Southampton

The inspection by Ofsted of courses of initial teacher education is high-stakes. An unsatisfactory report can lead to course closure. Even a satisfactory report can lead to reductions in quota resulting in a spiral of decline in course viability. The high-stakes nature of the inspection means that there has to be complete confidence in the level of validity and reliability of the inspection process. This paper presents an analysis of the complete cohort of published inspection reports of providers of secondary mathematics initial teacher education Postgraduate Certificate of Education (PGCE) courses carried out by Ofsted in the period 1996/8. The analysis demonstrates that there is considerable variation in the reports, in terms of word length, how particular criteria seem to be applied, and how judgements are expressed. With the complexity of the framework for inspection, it is impossible, given the current model of inspection report, to properly distinguish between consistency of application and the loading given to any particular criterion. Attention to the transparency of the inspection process and to matters of validity and reliability is crucial if there is to be confidence in the inspection system.

No measurement process is precise. All measurements have associated errors. While these realities are well-known in educational research (see, for example, Linn 1989), such matters have recently been the subject of papers from members of the (UK) Qualifications and Curriculum Authority (QCA) and of the Office for Standards in Education (Ofsted). For the QCA, the particular concern is the validity of national curriculum assessment (see, Stobart 1999); for Ofsted, the matter is the reliability and validity of school inspection judgements of (school) teaching quality (see Matthews *et al*, 1998). In this paper our concern is the validity and reliability of the process of judging the quality of courses of initial teacher education (ITE). For the purposes of analysis we focus on secondary mathematics PGCE courses in England, although the methodological approach we adopt, and the issues we raise, are likely to apply to other such courses inspected under the same framework.

The validity and reliability of the Ofsted inspection of ITE has been the subject of some inquiry already. For example, in a survey of providers of ITE courses, Graham and Nabb (1999) found that less than one in ten of 152 providers were confident that the inspection of ITE courses was a valid, reliable and consistent process. The survey also pointed to the use by Ofsted of 'additional inspectors' (AIs), people recruited and trained quickly alongside the process of inspection using HMI exemplification material, as being particularly problematic because they were judged to be even more inconsistent than the permanent inspectorial staff of Ofsted. Campbell and Husbands (2000) raise similar concerns and conclude, from their case study of the inspection of one particular institution, that 'the methodology of inspection is .... insufficiently reliable for the consequences which flow from it'. The comprehensive review of Ofsted recently completed by the Education Sub-committee of the House of Commons Select Committee on Education and Employment (House of Commons Select Committee on Education and Employment, 1999) also commented on the need to establish the level of reliability and validity of the basic elements of inspection. The committee expressed the wish to see research into this issue extended. It is important, they say, to help ensure 'public acceptance of inspection, that such work is open to scrutiny by the academic community' (*ibid.* para 129).

The full technical process of investigating the validity and reliability of the evaluation or measurement of educational endeavours requires access to the original data collected in order to make the judgements. To date, Ofsted have been disinclined to publish or allow access to such original data. What are available are the published reports on the inspection of each ITE provider. It is these public documents that we subject to analysis in order to see what they may reveal about the validity and reliability of the ITE inspection process.

As there are considerable variations in arrangements between the different constituents that make up the UK, we have chosen to focus on ITE providers of secondary mathematics PGCE courses in England. Building on our previous work in this area (Jones and Sinkinson 2000), we examine whether or not it is possible to properly distinguish between consistency of application of the published inspection framework and the loading given to any particular criterion within the framework.

The Inspection Framework for Courses of Initial Teacher Education

The UK Government inspectorate has been involved in inspecting teacher education institutions since the middle of the 19th century. In 1994, the newly constituted Ofsted was given expanded functions with a statutory duty to inspect ITE in all education departments in universities and other higher education institutions. In 1996 Ofsted embarked on a programme of inspecting all secondary school level courses of ITE, completing this in 1998. By the end of 1999, inspection reports were available covering every provider of secondary mathematics PGCE. It is these reports that serve as the data for the research reported in this paper.

Since 1996, the inspection of courses of initial teacher education in England has been determined by a Framework for the Assessment of Quality and Standards in ITT produced by Ofsted, in consultation with the (UK) Teacher Training Agency. The first version, published in 1996 (Ofsted/TTA, 1996), was used for the inspections carried out in 1996/7. This framework was then revised (Ofsted/TTA, 1997) and this new version used for the remaining inspections carried out during 1997/8. The framework (in both versions) requires the Ofsted inspectors to organise their judgements about the quality of ITE courses around a series of 'cells', each cell being graded on a 1–4 scale and each cell underpinned by a series of criteria. Grade 1 signifies 'very good, with several outstanding features', grade 2 'good, with no significant weaknesses', grade 3 'adequate, but requires significant improvement', and grade 4 is 'poor quality'. The 1997/98 version of the framework contains approximately 160 criteria statements for judging an ITE course. Not all these were used as Ofsted and the TTA announced that some 'cells' would not be considered during that round of inspections. Thus, during the inspections of secondary PGCE courses around 110 criteria were used.

One cell which was part of the inspection process was cell T2 (the quality of the training process). This cell has nine criteria, an example being 'training sessions exemplify good teaching'. Eight of them were graded by the inspector. Once each criterion has been graded by the inspector, the inspection guidance indicates how the overall cell grading is to be made. For example, to receive a grade 1 for any cell, 'most criteria will be judged to be very good and none less than good. There will be only a few of the criteria judged to be good (indicative range 20%-30% depending on significance)'(Ofsted 1996, 1997). More details of the inspection framework are given in Appendix A.

For provision to be judged compliant with the requirements of the Secretary of State for Education, each cell must be judged by the inspection process to be at least adequate (a grade 3). Without compliance in *each* of the cells, in *all* of the ITE courses that the institution provides, accreditation of the institution by the TTA is withdrawn. Further, on the basis of the courses they run, providers of ITE are rated on an A-E scale, where A is the highest grade. The TTA then uses these grades for accreditation purposes and to inform decisions on the allocation of places for ITE courses. Thus there is a close link between the outcome of the inspection of any course and the viability and reputation of the ITE provider. This 'high stakes' nature of the inspection process means that it is all the more vital to determine the validity and reliability of the inspection process. Unsatisfactory inspection reports have already led to the closure of courses and, in some cases, whole institutions (see, for example, Ghouri and Barnard 1998).

The Validity and Reliability of the ITE Inspection Process

Validity and reliability are central concerns for any process of educational measurement. While validity has traditionally been defined as the extent to which a 'test' measures what it purports to measure (for example, Cronbach 1949, p. 48) more recently the definition has been expanded. The traditional concept of validity considered content, construct, and criterion as three major but separate aspects of validity. This concept has increasingly come to be regarded as fragmented and incomplete, failing to take into account evidence of value implications of both score meaning as a basis for action and the social consequences of score use. Thus Messick (1989a, p. 13) states that validity is an 'integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment'.

Reliability refers to the accuracy (that is, the consistency and stability) of measurement by the measuring process (see Feldt and Brennan 1989). For example, the American Educational Research Association (1982, p.

1589) suggest that reliability 'concerns the extent to which measurements are repeatable, that is, when different persons make the measurements on different occasions, with supposedly alternative instruments for measuring the same thing'. Campbell and Husbands (2000) offer a definition of reliability of ITE inspection in England based on their knowledge of ITE and its inspection. The definition they offer is that, 'the criteria in the inspection methodologies and the Framework are applied consistently on inspection of different subjects and on different inspections of the same subjects within an institution and across institutions'.

For the research reported in this paper we employ the wider definition of validity (from Messick) and the definition of reliability suggested by Campbell and Husbands, given above. The issues of validity and reliability in the inspection process have prompted Ofsted inspectors to investigate the matter, at least as far as school inspections are concerned. Matthews *et al* (1998), report on the judgements of 100 pairs of trained school inspectors whom independently observed the same school lesson. Agreement occurred in 80% of cases. In 3% of cases, the judgement differed by as much as two grades. Matthews *et al* found 'less agreement on grades .. at a sensitive area of the grading scale' (*ibid.*, p.184), the boundary between satisfactory and unsatisfactory. They also found more agreement about *weaknesses* than about strengths and suggest a problem with the school inspection procedure is that 'inspectors take more care to record all the weaknesses .. than they do to record all the strengths'. They conclude that 'the anomalies here call into question the reliability of the judgements made by a very small proportion of inspectors on one occasion' (*ibid.*, p.186). Given that inspections of secondary subject courses of initial teacher education are generally carried out by a single inspector, even one instance of unreliability could have devastating results. This underlines the important of investigating the validity and reliability of the ITE inspection process.

Methodology

The main methodological approach adopted in this study of published Ofsted reports is one of critical document analysis. As Jupp (1996 p.311) explains, this involves 'a critical reading of texts aimed at uncovering how problems are defined, what explanations are put forward and what is seen as the preferred solution. It also seeks to bring to the surface that which is rejected in the text and that which does not even appear: what is *not* seen as problematic, what explanations are not considered, and what are not preferred solutions'.

We also utilised the 'criteria for assessing the quality and standard of inspections and the work of (school) inspectors' as published by Ofsted (Ofsted 2000). The specific criteria involved were:

- evidence is sufficient in quantity and range to be representative
- there is careful analysis and interpretation of all inspection information
- judgements are fully consistent with the inspection evidence, reflect reliable use of the criteria in the Framework
- the judgements made cover the relevant requirements in the Framework

The available literature noted above suggests that, in critically examining the validity and reliability of the Ofsted reports, we need to look for evidence of the demarcation of grade boundaries, particularly what distinguishes one grade from another, examine the tone of reports to see whether descriptions of weaknesses outweigh expressions of course strength, and investigate the reliability of the judgements, by searching for possible inconsistencies across equal grades.

We note that Ofsted reports are Crown Copyright and that a condition of use is that 'extracts quoted are reproduced verbatim without adaptation and on condition that the source and date thereof are stated'. In reporting our analysis we are conscious of the ethical issues involved in identifying individual institutions. All the extracts we reproduce below are quoted verbatim without adaptation from the inspection reports published by Ofsted as a result of the round of inspections of secondary mathematics PGCE courses carried out during 1996/7 and 1997/8 (Ofsted, 1998/99).

Results

We have analysed a total of 64 Ofsted reports on secondary mathematics ITE courses. Of the inspections carried out in 1996/7, 20 of the providers inspected were partnership schemes run by HEIs; one was a school-based scheme or SCITT. Of the 1997/8 inspections, 38 providers were HEI partnerships, with 5 being SCITTs. All the inspections carried out in the 1996/7 round were, we believe, carried out by HMI – inspectors employed and trained by Ofsted. There is considerable evidence that this was not the case for the completing round of inspections carried out in 1997/8 when the majority of inspections were done by 'additional inspectors' selected by Ofsted as being suitable for the job and given a modicum of training

amounting to two or three days (see evidence provided to the inquiry of the Education Sub-committee of the House of Commons Select Committee on Education and Employment (House of Commons Select Committee on Education and Employment, 1999)).

Given the definitions of validity and reliability adopted for this research, there is clearly a relationship between the two issues. The wider approach to validity subsumes considerations about reliability, which traditionally has been treated as a separate construct. This subsuming is because confidence in any inferences must include confidence in the results themselves. Any inconsistencies in the administration or grading undermines these inferences. In order to overcome validity and reliability being seen as two competing constructs there is an increasing preference for the use of dependability as a term which embraces both, what Gipps (1994) refers to as 'the intersection of validity and reliability'. In what follows we first consider issues to do with the framework for inspection. We then go on to consider the results of our analysis of the inspection reports. As we will demonstrate these results impinge on both the validity and reliability of the ITE inspection process.

The framework for inspection of ITE

Adapting Gilroy and Wilcox (1997) we can highlight three assumptions behind the Ofsted ITE inspection framework. These are, first, that the criteria are generally accepted as defining good ITE, second, that the meanings of the criteria are immediate and transparent, and third, that the application of the criteria is a straightforward process. Gilroy and Wilcox find major problems with all three assumptions. For example, while in adopting the framework, it was widely circulated for consultation, such circulation is unlikely to resolve differing views. As the evidence of the Open University to the Education Sub-committee of the House of Commons Select Committee on Education and Employment (House of Commons Select Committee on Education and Employment, 1999, appendix 76) makes clear, in their view the inspection framework is based on a limited conception of quality. First, they say, a narrowly defined orthodoxy of what is appropriate in ITE is developing. Second, continual re-inspection, in which providers are required to focus resources on narrowly defined issues of compliance, threatens development and innovation. Third, the omission of an analysis of value-added by ITE courses from the inspection framework means that no acknowledgement is made of the effectiveness of providers who take applicants from threshold entry level through to competent outcome levels.

The above points to some concerns about the validity of the inspection framework. In addition, that the framework has been changed puts some question on its validity. We now turn to our analysis of the published inspection reports.

Analysis of the published ITE inspection reports

Table 1, below, shows the distribution of grades given in the 1996/7 and 1997/8 secondary mathematics inspections.

Cell	Grade	1		2		3		4	
		96-97	97-98	96-97	97-98	96-97	97-98	96-97	97-98
S1	selection procedures	33%	19%	52%	63%	14%	19%	0	0
T2	quality of training	24%	21%	43%	60%	33%	19%	0	0
T4	assessment of trainee teachers	19%	23%	57%	51%	19%	21%	5%	5%
ST1	trainee teachers' subject knowledge	33%	12%	57%	77%	10%	7%	0	5%
ST2	trainee teachers' planning and teaching	24%	5%	48%	72%	29%	23%	0	0
ST3	trainee teachers' assessment of pupils	14%	5%	62%	74%	24%	16%	0	5%

Table 1: Grade profile of secondary mathematics PGCE courses,
1996-97 inspections (n=21, 20 HEIs, 1 SCITT)
1997-98 inspections (n=43, 38 HEIs, 5 SCITTs)
(Rounding errors account for the fact that not all totals are 100%)

A cursory examination of the distribution of grades appears to suggest that was more difficult to be awarded a grade 1 in the 1997/98 inspections compared to the 1996/7 inspections. It is worth noting that all the grade 4s in 1997/98 were awarded to SCITTs. Some possible reasons for such a difference may be:

- simply coincidence that there were more outstanding aspects seen in 1996/97
- the fact that the framework for inspection changed between the two years may have had an effect
- there was a marked difference in the personnel carrying out the inspections in the two years. In 1996/97 we believe that all inspections were conducted by HMI, in 1997/98 approximately two-thirds of the inspections were conducted by Additional Inspectors (AIs)

All three of these possibilities will be explored in more detail later in the paper.

As in our analysis of the 1996/97 inspection reports (Jones and Sinkinson 1999, 2000), we focused on cells T2 (quality of training) and C2/ST2 (trainee teachers' planning, teaching and classroom management).

In 1997/98 cell T2 (quality of training) comprised nine criteria, of which eight were to be judged by the inspector. Hence a provider might reasonably expect each of the eight criteria to receive a specific mention in the final inspection report. Of the 43 reports scrutinised for this part of the research, 12 [28%] made no specific reference to criterion f: 'the training is differentiated to build on trainees' academic background and relevant experience'. This compares with a similar omission in approximately 60% of reports in the 1996/97 round, so perhaps the situation is improving.

The length of the report on each cell varies considerably, as it did in the 1996/97 round; here the variation within cell T2 is from 4 to 10 paragraphs, with most paragraphs reporting on more than one criterion within the cell. Given that we are forced into a judgement system which lists exactly what inspectors wish to see, perhaps it would be an advantage if the method of reporting was standardised so that each criterion was afforded one paragraph in which the exemplar evidence was listed alongside the judgement for that criterion. In only just over half of the reports [23/43, 53%] was any evidence cited alongside particular judgements; for example 'training sessions in schools are mostly excellent, based on concise observation, specific and constructive feedback, and rooted in target setting and evaluation', 'In a very good quality tutorial given by an associate tutor, the trainee confidently led a discussion to establish a number of strategies for effective marking of pupils' work.' 'An outstanding feature is that all components [of the PGCE course], including school-based tasks and assignments, are integrated well and each is underpinned by good documentation'. It is, in the authors' opinion, quite difficult, on the basis of the number of school-based training sessions which inspectors see at each phase of the inspection, to make such a generalisation as that given in the first quote but, by including such exemplars within the reports inspectors are, at least, giving providers some insight into what inspectors consider to be aspects of good practice.

Some examples of possible inconsistencies and anomalies found in cell T2 within the inspection reports for providers inspected in 1997/98 follow. Whilst we recognise that these statements are only sections of the complete cell report, such inconsistencies can be seen within complete cell reports which have been awarded different grades.

> Feedback given to the trainees following observation of their teaching is mostly good, and at times is excellent [Grade 2]
>
> They (mentors) give good quality oral feedback immediately after lessons [Grade 1]
>
> Whilst no subject profile or rigorous audit exists, the partnership successfully assesses and monitors subject knowledge. Mentors and university tutors provide effective support [Grade 1]
>
> Assessment of the trainees' subject knowledge is mostly done at interview, and is continued informally by the college tutors during mathematics topic training sessions. Trainees also use a subject knowledge computer based training package, according to their needs [Grade 2]

Thus there remains some concern about borderline judgements of criteria in particular.

Nine providers were awarded a grade 1 in cell T2 in 1997/98. As with the 1996/97 round, there appears to be considerable disparity in the descriptors used to validate these judgements. A Grade 1 is defined as 'very good with several outstanding features,' (Ofsted and TTA, 1997, p7). The spread of descriptors used within each cell awarded a Grade 1 is given in Table 2.

Institution number	Number of statements expressing very good or equivalent	Number of features listed as outstanding, excellent, or exemplary
1	2	1
2	2	3
3	4	1
4	3	8
5	7	2
6	6	2
7	1	3
8	1	3
9	7	1

Table 2: Analysis of reports awarding a Grade 1 in cell T2
(Grade 1 means 'very good with several outstanding features')

In contrast, the profile for several of those providers who were awarded a Grade 2 in cell T2, is shown in Table 3.

Institution number	Number of statements expressing very good or equivalent	Number of features listed as outstanding, excellent, or exemplary
1	1	1
2	1	2
3	0	1
4	2	1
5	3	2

Table 3: Analysis of some reports awarded a Grade 2 in cell T2
(Grade 2 means ' good with no significant weaknesses')

Although there seemed to be some anomalies, particularly at the actual boundaries between grades 1 and 2, there appeared to be a clearer distinction in the overall cell reports between providers graded 1 and those graded 2 in cell T2, than was evident within the 1996/97 reports (Jones and Sinkinson, 1999, 2000). This may indicate that internal moderation between inspectors is becoming more solid or that they have found the 1997/98 framework for inspection easier and more effective with which to work.

As stated earlier, the Framework for Inspection was changed in between the 1996/97 inspections and those conducted in 1997/98, even though they were all part of the same 'round' of inspection. The Secretary of State's requirements and criteria for courses in initial teacher training are given in DfEE Circular 10/97, in which the changes between the two frameworks are summarised thus:

> A major change within this Framework is the replacement of what were previously described as competences (in the C cells) with the new QTS standards (now in ST cells)...In other respects, partly in response to informal feedback from providers, this version remains very similar to the version developed for 1996/97.
>
> (Ofsted and TTA, 1997, p1)

In terms of cell T2 (quality of training) this is indeed true, the nine criteria which comprise this cell are identical in each edition of the framework. The major differences occur in the content of cell ST2 (previously C2) and in the strategies for allocating individual cell grades. In cell ST2 there are 14 criteria, one of which is divided into five subsections and another into fourteen subsections. In essence, ST2 has become a collection of 31 individual criteria, almost three times the 11 criteria inspected for the equivalent cell, C2, in 1996/97. It may be that the sheer number of criteria within ST2 accounts, at least in part, for the apparent disparity in proportions of providers being awarded a Grade 1 in cell C2/ST2 in each of the two years; namely 28% in

1996/97 and 6% in 1997/98, see Table 1. However, not all 31 criteria are graded: criterion ST2k has 14 subcriteria each of which, on inspection, appears likely to be as important as any other criterion within ST2, yet only one grade is awarded which, we assume, is an amalgamation of all evidence collected across all 14 subcriteria. The same is true for criteria ST2a, which has five subcriteria. Hence, we concur with the quotation above that there is a major difference between the two Frameworks used within the same inspection cycle.

The variation in the lengths of reports written on different cells is repeated in the 1997/98 inspections: in cell ST2 the range is from 3 to 10 paragraphs, with the median being 6. As in 1996/97, there appears to be no obvious relationship between the length of the report on ST2 and the grade awarded. The differences in the ways in which the cell reports are constructed present problems of inconsistency, particularly within what is designed as a criterion based evidence base. As Campbell and Husbands (2000, p46) suggest, there appears to be a disparity between the model of inspection, in which the criteria are, according to Ofsted, set out in clear and unambiguously accountable terms, and the content of the actual, published reports of each inspection. It seems reasonable to suggest that, since not all criteria within ST2 are mentioned specifically in every report, at least some of the judgements for some providers were arrived at through what Campbell and Husbands (*ibid*.) describe as 'a rather more traditional HMI model.' Such a model relies less on a formal checklist of criteria and more on inspectors making a judgement of individual cell quality which fits in with their evidence-based judgement about the overall coherence and quality of the inspected course.

Issues of reliability are paramount within this approach. It is likely that most HMIs are able to do this effectively and reliably, but the authors are less confident that AIs, given the length of training they receive prior to inspecting, have the 'tacit HMI expertise' described by Campbell and Husbands (*ibid.*). Although the authors of this study chose not to seek data about the 'type' of inspector conducting each inspection, it may be a useful line of further enquiry, particularly as Campbell and Husbands (1999) have published evidence of overtly unreliable judgements by AIs; it would be reasonable to assume that there are other such instances. Indeed, Graham and Nabb (1999, p24) report that 59% of the respondents to their questionnaire felt that judgements made by AIs were less valid and reliable than those made by HMIs.

As stated earlier, the frameworks under which courses were inspected in the two years were actually quite different, yet the results have been combined and reported on as if they were all conducted under the same set of rules. The 1996/97 inspections were conducted by HMI whilst approximately two-thirds of the 1997/98 inspections were carried out by Additional Inspectors (AIs), appointed by Ofsted solely for the purpose of these ITT inspections. We understand that training was restricted to a three day course followed by some 'shadowing' with an HMI, but it was not a precondition of appointment that applicants should have personal experience of being a teacher educator.

We have shown earlier that the distribution of grades, particularly with respect to grades 1 and 2, is quite different in 1997/8 from that in 1996/97; it appeared to be much harder to be awarded a grade 1 in 1997/98. Perhaps this has something to do with the fact that so many inspections were undertaken by AIs in 1997/98. However, it may be related to the differences in the recording of evidence procedures and the methods by which inspectors arrived at a summative grade for each cell. Although Ofsted have not published accounts of how individual cell grades were obtained, details available to providers in the 1996/97 round of inspections, (Ofsted, 1996), show that each individual criterion within a cell was graded by the inspector, who then calculated a final cell grade based upon those interim grades. The situation in the 1997/98 inspections is described above, showing that the grading of many subcriteria was subsumed within a criterion grade, despite it appearing that subcriteria are as important as individual criteria, (see Ofsted 1997).

Part of the evidence collected by inspectors for cell ST2 is lesson observation of trainees towards the end of their final assessed teaching placement. These observations are graded both on Ofsted's Guidance on the Inspection of Schools grading system, which is a seven point scale. In addition, each observed trainee is awarded a grade 1 - 4 or 9 for each of cell C1 - 3 or ST1 - 3, based upon the grades for individual criteria within the cell, as described above. The overall cell grade for each trainee observed teaching 'for approximately one hour', (Ofsted 1996, 1997), is determined by the inspector. Neither the two Frameworks for Inspection nor the guidance materials (*ibid*.) specify what strategies were employed to arrive at the final cell grading. Under a section on 'grading' within the Secondary Subject Inspections Guidance (*ibid*.) we learn that, for the award of a grade 1 'several outstanding features will need to be identified and most criteria will be judged to be very good and none less than good. There will be only be a few of the criteria judged to be good (indicative range 20% to 30% depending on significance)'(Ofsted 1997). The interpretation of the word

'significance' appears to be left to the individual, the authors could find no exemplification within any inspection documentation for either year. Similar difficulties arise with interpretation of the term 'several' and these have been shown to lead to the inconsistencies illustrated in Table 2. Such vague definitions do seem to leave a great deal of potential for lack of reliability and validity such as that described by other researchers within this area, for example Graham (1997, p6), Campbell and Husbands, (2000) and Jones and Sinkinson (1999, 2000). The way in which criteria are interpreted, the way in which some or all of the criteria listed within each cell are combined to form judgements and eventual cell grades are all significant factors affecting reliability; at present there is insufficient detail forthcoming from Ofsted in all of these issues to promote confidence among providers of secondary mathematics PGCE courses. Similar apparent inconsistencies over statements within the cell reports on ST2 are evident; we focus here on examples which appear to be targeted against criteria within a) the planning section of ST2 and b) the classroom management section.

> Trainees plan sequences of lessons well drawing on the appropriate GCSE syllabus or national Curriculum Programme of Study and make very good use of assessment information to ensure that lessons are challenging [Grade 2]
> Trainees plan their teaching well to achieve progression in pupils' learning. Good use is made of suitably challenging tasks and trainees plan well for pupils to be motivated and interested by the content of lessons [Grade 1]
> All trainees manage their classes well [Grade 2]
> Classes are generally well managed. trainees expect and, in the main, uphold high standards of behaviour from pupils, maintaining a purposeful working atmosphere throughout the lesson. Most trainees deal effectively with misbehaviour without loss of good humour and maintain productive relationships with classes [Grade 3]
> Trainees teach whole classes, groups and individual pupils effectively. Most respond firmly to pupils' misconduct, reproaching without confrontation, and so maintaining order with good humour [Grade 2]
> Trainees manage pupils well, maintaining a good standard of discipline and a positive working environment [Grade 1]

An additional, apparent inconsistency which, we suggest, may have a bearing on the reliability and validity of published data concerning inspection outcomes, is evident within the 'Performance Profiles' through which the TTA publishes inspection grades for each provider, (TTA, 1999). It concerns the gradings given to Key Stage 2/3 Mathematics PGCE courses, which are deemed to be secondary courses for the purpose of numbers allocations to providers. To the authors' knowledge, no KS 2/3 PGCE Mathematics course has been inspected formally by Ofsted to date. The grades awarded within this publication are, in every case, identical to those awarded to the provider's secondary mathematics course. Further, the TTA states that 'secondary inspection evidence is also used to determine the quality category of Key Stage 2/3 courses' (TTA 1999). We contend that this is neither reliable, valid, nor based on any evidence whatever. There is no generalisable evidence to suggest that there are any specific links or relationships between content, partnership schools, quality of training or of trainees, within the KS 2/3 PGCE Mathematics courses and the secondary ones to which the gradings are tied.

Discussion

It is abundantly clear that education, at all levels, is now rooted firmly within what Norris (1998) calls an 'evaluation culture.' The procedures involved in such evaluations have been shown, in this research and in that conducted by others, for example Campbell and Husbands (2000), to be lacking in terms of reliability. Issues concerning funding allocations, trainee numbers and institutional reputations, not to mention lecturers' jobs, are a direct consequence of the outcomes of inspection. hence it is vitally important that all involved in the inspection of PGCE courses have confidence in both the methodology adopted and the judgements made.

Campbell and Husbands (*ibid.*) argue that it is difficult to establish how far Ofsted has succeeded in addressing problems of reliability such as 'inter-observer agreement, representativeness of classroom behaviour sampled and the influence of observers upon the observed', since Ofsted has not, to date, allowed its policy and practice in these areas to become part of the public domain. Gilroy and Wilcox (1997) have shown the virtual inevitability of inspection criteria being specified imprecisely - such variability has been described earlier in this paper and illustrated through reference to statements taken from inspection reports. Campbell and Husbands (2000) provide further evidence of the potential for inconsistency of interpretation of criteria.

In this analysis we have relied heavily upon Campbell and Husbands' definition for testing reliability, namely: the criteria in the inspection methodologies and the Framework are applied consistently on inspections of different subjects and on different inspections of the same subjects within an institution and across institutions', (Campbell and Husbands, 2000, p42), and have provided several examples of inconsistencies which we suggest, affect reliability adversely. Evidence from Graham and Nabb (1999, p22) suggests that 79% of ITT providers who responded to their questionnaire are not confident that ITT inspection is a valid, reliable and consistent process, so there is clearly a problem, if only a perceived problem.

Ofsted argue that national moderation meetings and exemplification details of criteria provided during training for inspectors deal adequately with potential difficulties of inconsistency. this may indeed be true, but the procedures adopted by Ofsted are kept behind Ofsted's walls. The rules governing the moderation remain similarly secret, as do the grades awarded to individual trainees and to individual criteria within each cell. Thus, it is virtually impossible for any provider to monitor consistency. It would be a huge step forward in terms of providers striving to constantly improve the training that they offer, for the exemplification details to be placed within the public domain. This might give substantial credibility to Ofsted's stated aim of 'stimulating and informing discussion and contributing to the development of policy and practice in secondary initial teacher training'. (Ofsted, 1999, p6).

At present there appear to be little confidence amongst providers that the feedback provided by Ofsted does indeed contribute to the development of practice. Graham and Nabb (1999, p20) report that 75% of ITT respondents disagreed or strongly disagreed with the statement: 'The ITT system receives sufficient overall feedback about good practice based upon inspection evidence'. This seems to indicate that providers would welcome more statements, within their inspection reports, which provide examples of good practice seen during the inspection.

This research indicates that there is room for much development in order that all participants in the process of inspecting secondary mathematics PGCE courses are confident that it is reliable, valid and robust. Perhaps there is a need for a discussion between Ofsted and providers about the merits of the 'traditional HMI approach to inspection' which existed prior to 1996, in which there appeared to be much more confidence, and the criterion-based approach in place now, which is th source of so much disquiet.

Acknowledgements
Our work in preparation for this paper has benefited from the feedback we received from colleagues at the 4th British Congress on Mathematical Education and the 1999 conference of the British Educational Research Association. We are grateful for the encouragement we have received from these colleagues and for the interest shown in this work by members of the Ofsted inspectorate.

References

American Educational Research Association, (1982), *Encyclopaedia of Education Research*, vol. IV. New York: Free Press.

Campbell, J. and Husbands, C (2000) On the Reliability of Ofsted Inspection of Initial Teacher Training: a case study. *British Educational Research Journal*, **62**(1), 39-48.

Cronbach, L. J. (1949). *Essentials of Psychological Testing*. New York: Harper & Row.

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

Gilroy, P. and Wilcox, B. (1997) Ofsted, Criteria and the Nature of Social Understanding: a Wittgensteinian critique of the practice of educational judgement. *British Journal of Educational Studies* 45(1), 22-38.

Gipps, C.V. (1994), *Beyond Testing: toward a theory of educational assessment*. London: Falmer Press.

Graham, J. (1997) A Quality Experience? The TTA/Ofsted quality framework in initial teacher education 1996/7. In *Initial Teacher Education: TTA/Ofsted quality framework.* London: UCET.

Graham, J. and Nabb, J. (1999), *Stakeholder Satisfaction: survey of Ofsted inspection of ITT 1994-1999. UCET Research Paper No. 1.* London: Universities Coucil for the Education of Teachers.

Ghouri, N. and Barnard, N, (1998), Training Courses go after Poor Inspections, *Times Education Supplement*, October 30 1998.

House of Commons Select Committee on Education and Employment (1999) *The Work of Ofsted: 4th report of the education and employment committee, H.C. papers 62-I, session 1998-1999*. London: The Stationery Office.

Jones, K. and Sinkinson, A. (1999), *Analysis of Ofsted Judgements in Secondary Mathematics PGCE Courses, 1996-97*. Paper presented at the British Educational Research Association Annual Conference, University of Sussex.

Jones, K. and Sinkinson, A. (2000), A Critical Analysis of Ofsted Judgements of the Quality of Secondary Mathematics Initial Teacher Education Courses. *Evaluation and Research in Education*.

Jupp, V. (1996) Documents and Critical Research. In R. Sapsford and W. Jupp (Eds), *Data Collection and Analysis*. London: Sage.

Linn, R. L. (Ed.) *Educational Measurement*. Washington, DC: American Council on Education/Macmillan. 3rd edition.

Matthews, P., Holmes, J. R., Vickers, P. and Corporaal, B. (1998) Aspects of the Reliability and Validity of School Inspection Judgements of Teaching Quality. *Educational Research and Evaluation* **4**(2), 167-188.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational Measurement*. Washington, DC: American Council on Education/Macmillan. 3rd edition. (pp. 13-103).

Norris, N. (1998) Curriculum Evaluation Revisitied, *Cambridge Journal of Education*, 28, 207-220.

Office for Standards in Education/Teacher Training Agency (1996*)*, *Framework for the Assessment of Quality and Standards in Initial Teacher Training*. London: HMSO.

Office for Standards in Education/Teacher Training Agency (1997), *Framework for the Assessment of Quality and Standards in Initial Teacher Training*. London: HMSO.

Office for Standards in Education (1996), *Secondary ITT Subject Inspections 1996-97 Guidance.* London: HMSO.

Office for Standards in Education (1997), *Secondary ITT Subject Inspections 1997-98: Guidance.* London: HMSO.

Office for Standards in Education (1998 and 9), *Various Reports resulting from the round of inspections of secondary mathematics PGCE Courses carried out during 1996/97 and 1998/9*. London: HMSO.

Office for Standards in Education (1999),*Secondary Initial Teacher Training. Secondary Subject Inspections 1996-98 Overview report.* London: HMSO.

Office for Standards in Education (2000), *Criteria for Assessing the Quality and Standard of Inspections and the Work of Inspectors.* Ofsted Update for Inspectors, Issue 32, Spring 2000.

Stobart , G. (1999), *The Validity of National Curriculum Assessment*. Paper presented at the British Educational Research Association Annual Conference, University of Sussex at Brighton, September 2 - 5 1999.

Teacher Training Agency (TTA) (1999), *Initial teacher Training Performance Profiles: September 1999*. London: Teacher Training Agency.

Appendix A

The Framework for the Assessment of Quality and Standards in Initial Teacher Training

In both the 1996/97 and the 1997/98 framework there are three 'leading' cells. In 1996/97 these were 'C' (competence) cells and in 1997/98 'ST' (standards) cells:
ST1 The quality of students' subject knowledge for teaching in the relevant age range
ST2 The quality of students' planning, teaching and class management
ST3 The quality of students' monitoring, assessment, recording, reporting and accountability

The 1997/8 framework has an additional 'leading' cell:
ST4 The trainees' knowledge and understanding of other professional requirements

There are then a series of 'contributory' cells relating to the quality of training (T-cells), selection of students (S-cells), quality of resources (R-cells) and management of ITT (M-cells).

In 1996/97 and 1997/98 the inspections reported on C1-3/ST1-3 and two of the T-cells:
T2 The quality of the training process in developing standards ST1–3
T4 The quality and consistency of the assessment of students' standards ST1–3

and on one S-cell
S1 The appropriateness of the admission policy and selection process