Consistency counts – or does it?

Teacher Education Advancement
Network Journal

Vol 9(1) pages

Rick Tynan, Andrea Mallaburn
Liverpool John Moores University

Abstract
All stake holders in competency based teacher training systems share an interest in the consistency of assessment outcomes and practice. Assessment data from more than 200 trainees participating in Initial Teacher Training/Education (ITT/E) programmes and partnerships at a Higher Education (HE) provider in the Northwest of England were analysed during the academic year 2014-15.

At four formal review points the overall teaching grades received by trainees were compared across five ITT/E programmes leading to Qualified Teacher Status (QTS). Several statistical approaches were employed and compared. All the methods indicated consistency of outcomes across the programmes for the final summative assessment.

Two statistical methods were used to investigate the strength of correlations between grades awarded for individual teaching standards and the trainees' overall teaching grades. Both demonstrated that all individual standards were positively correlated with overall teaching grades.

The second and qualitative phase of the study is ongoing and uses Q-Analysis to illuminate these initial findings by seeking to identify clusters of subjectivity amongst mentors and tutors when prioritising statements about assessment. It is too early to report any results from this phase.

Key Words
assessment; placements; mentors; tutors; standards; competencies; ITT/E; consistency; primary; secondary.

Context
Achieving and demonstrating consistency in both assessment outcomes and practice is of interest to both HE (Higher Education) and School based accreditors of QTS (Qualified Teacher Status). Ofsted (2015) (Office for Standards in Education) uses consistency across partnerships as a performance indicator in the inspection of ITT/E (Initial Teacher Training/Education) provision. We set up this study to test the assumption that statistical analysis of assessment outcomes supported by qualitative evidence of assessment procedures can be used to demonstrate consistency in these areas. The aim was to use quick and reliable analysis tools and apply them diagnostically throughout the year in order to redress any inconsistencies detected between programmes or assessment points.

There are reasons other than inspection for seeking to improve consistency. Our experiences across school/university partnerships indicate that the quality of mentoring and coaching relationships can be adversely affected when trainees perceive their assessments to be inaccurate or unfair. On the other hand, there is relatively little central guidance to help

Citation
Tynan, R., Mallaburn, A. (2017) 'Consistency counts – or does it?', *TEAN Journal*, 9(1), pp

assessors make objective and reliable judgements about their trainees' competencies with respect to the Teachers' Standards (Department for Education, 2011) in England.

Recommendation for QTS is currently based upon the assessment of teacher competencies described by eight teaching standards split into a number of sub-divisions together with a set of professional expectations (Department for Education, 2011). Individual standards and overall teaching are graded using a four-point scale: 1 (Outstanding), 2 (Good), 3 (Requires improvement) and 4 (Inadequate). Assessors are instructed to take into account trainee experience and stage of training and to adopt a holistic approach to sub-section criteria when reaching a judgement about the grade for an individual standard (Department for Education, 2011). The standard descriptors set out minimum expectations for performance but provide no indication of what is required for the award of grades 2 and 1.

A wide range of schools and HE providers involved in ITT/E in the Northwest of England have collaborated over time to apply general Ofsted descriptors for the assessment of final year trainees to the Teachers' Standards criteria for minimum performance (Department for Education, 2011). The result has been an individual trainee standards tracking document containing a set of performance descriptors for all sub-sections of the standards at every point on the four-point grading scale. To facilitate consistency across the HE provider's partnerships all assessors were expected to use the tracking document to reference their grading decisions.

Across the programmes involved in the study the subject mentors supervising the trainee teachers in school assessed them at three formative and one final summative assessment point. These corresponded to the completion of each phase of training (Figures 2 and 3). They awarded grades for each of the Teachers' Standards and collated these judgements to arrive at an overall teaching grade. Professional mentors moderated assessments made by different teachers within their school and school liaison tutors from the HE provider visited schools to conduct training and quality assure the mentoring and assessment processes. However, from our experience across the partnerships, despite this high level of professional, organisational and individual effort, assessment and grading continues to challenge new and experienced mentors and tutors. In turn, achieving and gathering evidence of consistency in assessment within and across multiple partnerships and programmes is a challenge for those with quality assurance roles. An obvious place to look for evidence was the assessments data and partnership documentation generated by trainees.

This paper reports the quantitative results from the first year of a mixed method, practitioner research investigation into the consistency of assessment outcomes across one North West of England HE provider's ITT/E programmes and partnerships. We also report on progress with data gathering for the second qualitative phase of the project. The project is on-going but early indications are that these approaches are worth pursuing.

Methodology and methods

This investigation is a practitioner led staff project. It links to local perceptions of issues and opportunities around consistency of assessment outcomes and practice for schools working in ITT/E partnership with a Northwest of England HE provider. The study evaluates the impact of interventions intended to improve consistency within and across partnerships and has the potential to become cyclic. As such it fits well with an action research model of investigation (Burton & Bartlett, 2009: 9).

The planned interventions were to:

- increase attendance and participation in Mentor Training by including a training element in all liaison visits and supplementing the HE provider programme of training meetings by on-site training in partner schools
- maximise consistency of assessment outcomes and practice by referencing all assessments to the minimum performance descriptors set down in the Teachers' Standards (Department for Education, 2011) and the criteria set down in the individual trainee standards tracking document
- increase the rigour of the final assessment process through longer, more structured triangulation meetings chaired by HE Tutors

The assessment data analysed statistically were drawn from five programmes at four formal review points in the academic year 2014-15. The programmes involved were Primary PGCE, Secondary PGCE (Postgraduate Certificate in Education), Secondary Salaried School Direct, Primary Education Honours degree with QTS (3 Year) and Primary Education Honours degree with QTS (4 Year). Non-Salaried School Direct trainees were grouped with core PGCE trainees. The number of trainees following each programme varied according to the quotas allowed and final uptake by applicants. The trial included three ways of analysing overall teaching grades at different review points across programmes and two ways of comparing overall grades to the grades for individual standards.

The statistical analyses used to compare overall teaching grades across programmes and assessment points were:

- The visual presentation of mean grades and their 5% confidence limits
- Single factor analysis of variance (ANOVA) on counts for grades across programmes
- Chi squared analysis on counts for grades across programmes

- The statistical analyses used to compare individual standard grades with overall teaching grades were:
- Pearson's correlation coefficient
- Spearman's Rank correlation coefficient

For the qualitative phase of the investigation, a concourse of around 40 statements concerning the assessment of trainees on school placement has been constructed from policy and course documentation. Mentors and tutors from across programmes and partnerships have been invited to participate in an anonymous on-line activity to place these in personal priority order for Q analysis (Brown, 1980, van Exel et al., 2005). This will identify clusters of subjectivity with respect to the concourse of statements amongst respondents. Participation will be anonymous and voluntary. Informed consent will be implied by completing and submitting the on-line activity.

The full project will collate the findings from both the quantitative and qualitative data analysis and, as such, constitutes a mixed methods study.

Interpreting the statistical conclusions

The methods trialled were a mixture of parametric and non-parametric statistical analysis. Parametric tests assume that data are distributed in a particular way whereas non-parametric

tests do not. If you know the distribution of your data, you can usually place more confidence in a conclusion reached using an appropriate parametric method of analysis. The parametric methods trialled were all designed for use with normally distributed data. This distribution is often seen when observations from large populations are presented graphically. Plotting the frequencies with which values occur results in a characteristic bell shaped curve that is symmetrical either side of the average value. Norman (2010) argued strongly that there is evidence that conclusions reached using parametric methods can be robust and valid even when the assumptions underlying their use are not in place. However, when data do not (or are not known to) conform to any distribution then a non-parametric approach may give more reliable and accurate conclusions.

Statistical analysis starts with the Null Hypothesis (H^0). This is always a neutral or cautious hypothesis e.g. there is no difference between the mean grades or numbers of grades awarded by the five programmes at a particular assessment point. This is accepted or rejected and the four possible results from a statistical test are shown in Figure 1. Statisticians prefer to reject H^0 but only because the probability of a wrong conclusion is quantified when this happens. The maximum risk of error allowed in statistical investigations is typically a probability of 0.05 or 5%.

		Null Hypothesis (H^0)	
		H^0 Valid	H^0 Invalid
Conclusion	Reject H^0	**Type 1 Error (5% chance of error)**	**Correct**
	Accept H^0	**Correct**	**Type 2 Error (error unknown)**

Figure 1. Statistical errors.

Using statistical tests ensured that objective conclusions were reached about consistency or inconsistency in the assessment data. All the statistical methods used to compare overall teaching grades awarded across the five ITT/E programmes were interpreted in the same way. Accepting H^0 indicated consistency in assessment outcomes and rejecting it demonstrated inconsistency. For the investigation of correlations H^0 was that there was no correlation between grades for individual teachers' standards and the overall teaching grades awarded. Rejecting H^0 with a positive correlation suggested an association between the standard and overall teaching performance in the minds of the assessor. However, a negative correlation or no correlation indicated grading decisions about individual standards and overall teaching that were inconsistent with the guidance and training the assessors had received.

Results

The use of means and 5% confidence limits to demonstrate assessment data pictorially (Figure 2.) was visual and easily understood. It demonstrated the progression in overall teaching grades throughout the year awarded across all the programmes. The number of assessments made within programmes varied across the assessment points as some students deferred their studies, returned to study or permanently left their course. The 5% confidence limits of the mean overall grades awarded by programmes at the end of each phase of training overlapped

except for one pair of programmes at the third formative assessment point (Figure 2.). This suggested that there was a high degree of consistency between programmes.

Single factor ANOVA indicated no differences in the distribution of grades across the five ITT/E programmes at any assessment point (Table 1) suggesting consistency between programmes.

Chi squared analysis demonstrated specific differences between the observed and calculated expected frequencies of overall teaching grades between programmes for the first three assessment points but not the last (Table 2).

Norman (2010) discussed and defended the parametric analysis of data derived from number scales similar to those used for our trainee teacher assessments. However, for our data there were differences in the conclusions reached by different methods and, without going into detailed mathematical and statistical arguments, the method that we had most confidence in was the Chi Squared analysis. This is a non-parametric method comparing observed numbers of grades awarded with expected numbers calculated using a contingency table. It indicated differences between individual programmes and the rest of the partnerships at the formative assessment points but consistency between all of them at the final summative assessment.

The correlation coefficient analysis compared the grades awarded for each individual teaching standard with the overall teaching grade at each assessment point for each programme. The correlation study did not indicate any 'rogue' standards (not positively associated with overall teaching grade) in any programme at any assessment point. All correlations were positive with a 5% or less chance of this conclusion being in error. This indicated consistency of outcomes across programmes at all assessment points. The results were statistically interesting because both the parametric and non-parametric methods gave similar results with identical conclusions in all cases.

Assessment points and phase of training

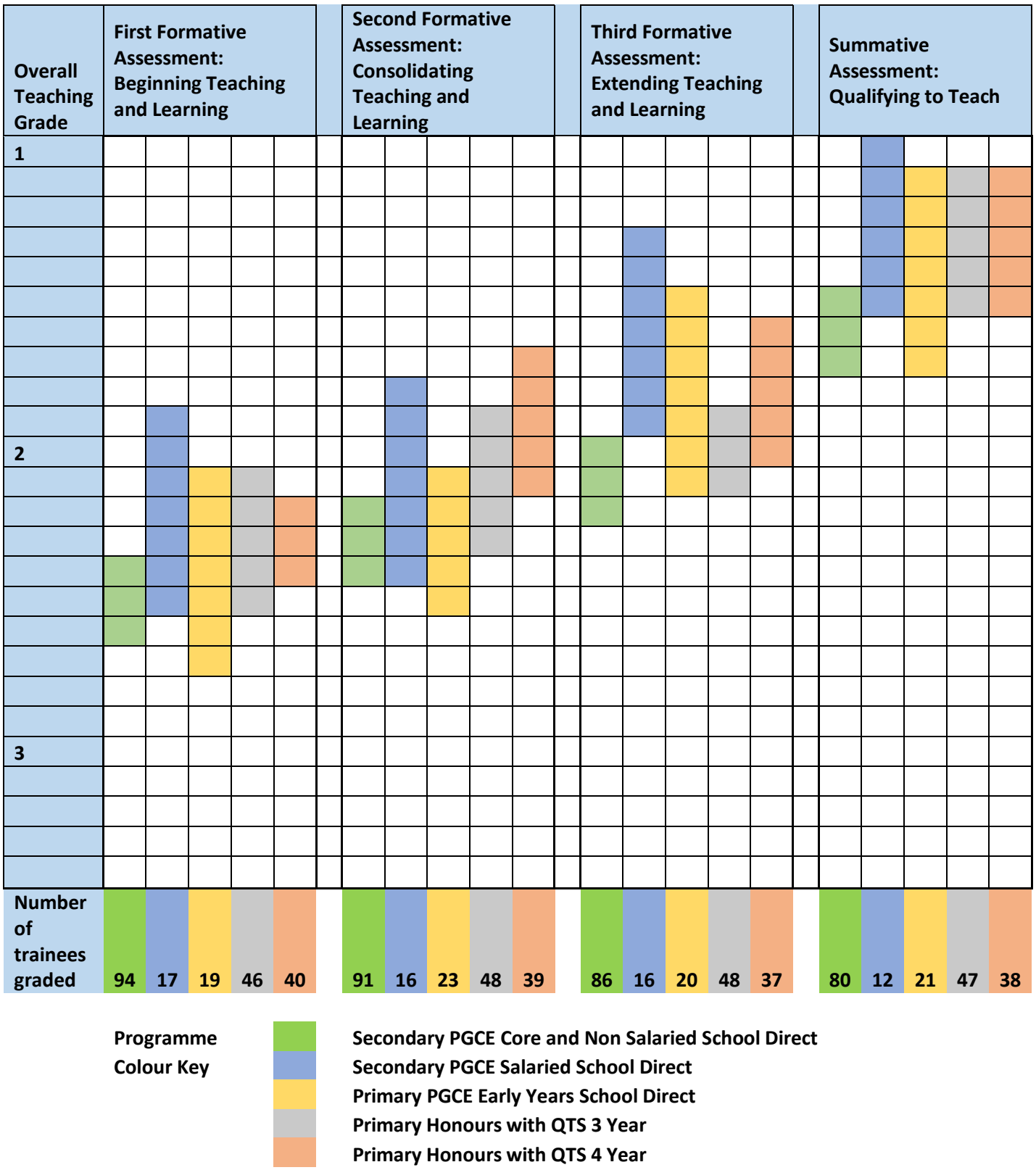

Figure 2. A pictorial representation of mean overall teaching grades and their 5% confidence limits across five programmes at each review point.

Table 1. ANOVA summary table.

Assessment	**F-value (the result of the ANOVA test)**	**Probability (P) of error if H^0 is rejected**	**Conclusion (reject H^0 if P is 0.05 or less)**
Summative	1.02	0.43	H^0 accepted
3rd Formative	0.64	0.64	H^0 accepted
2nd Formative	0.94	0.47	H^0 accepted
1st Formative	1.02	0.43	H^0 accepted

Table 2. Chi-squared summary table.

Assessment	**Degrees of freedom**	**Chi^2 value**	**Critical values Probability (P) of error if H^0 is rejected: 0.05 0.01**	**Conclusion (reject H^0 if the Chi^2 value is larger than either critical value)**
Summative	8	9.51	15.51 20.09	H^0 accepted
3rd Formative	4	15.55	9.49 13.28	H^0 rejected (P= 0.01)
2nd Formative	4	14.73	9.49 13.28	H^0 rejected (P= 0.01)
1st Formative	4	14.24	9.49 13.28	H^0 rejected (P= 0.01)

Discussion

Our study arose from our perception that the nature of professional learning and of criteria referenced assessment of competencies contribute strongly to the challenges associated with achieving and maintaining consistency in assessment across ITT/E programmes and partnerships. Philpott (2014) provided a critical summary of professional learning models and their relationship to the current school led model of ITT/E provision. He divided these broadly into those that focus on the psychology of individual learning and those that emphasise group or social aspects of learning. Kolb's experiential learning model and the clinical practice approach are examples of models that start with the individual's cognitive development and the increase in knowledge and skills based upon the evidence of practical experience (Philpott, 2014). Communities of practice and apprenticeship models emphasise social aspects of learning and the need for trainees to demonstrate independent competence to gain acceptance as a practitioner (Philpott, 2014).

Models emphasising the trainee's individual development of professional knowledge and skills appear to give more opportunity for assessors to be objective in their judgements. However, criterion referenced assessment of students in HE is liable to a variety of subjective influences even when assessments have been designed to reduce this (Donovan, Price and Rust, 2001). Further, the Teachers' Standards (Department for Education, 2011) describe the minimum criteria for competence but give no guidance on acceptable evidence to use in judging when a standard has been achieved or at which grade. Ofsted (2015) perceives successful teaching in terms of the pupils' learning outcomes. The argument is that a standard is met when its impact on learning is at least satisfactory over time. Learner rather than teacher performance then becomes the evidence for standards and the bigger and more consistent over time the impact

on learning the higher the grade awarded. However, our experience indicates that the problem of establishing how much a trainee's competency in a particular standard contributes directly to the overall impact on pupil learning appears to remain essentially subjective.

Learning to become a teacher is not a straightforward process that can be tracked using simple assessment tools. Hager and Butler (1996) proposed a model for professional development that also considered assessment. Martin & Cloke (2000) applied this model to teaching and the assessment of teacher competencies. Without troubling too much about the individual or social processes involved in professional learning their model highlights differences in activity and expectations as professional development progresses, and the assessment models associated with each stage. Figure 3 maps their model to the phases of teacher training currently in common use in ITT/E partnerships in the Northwest of England. We have found the judgemental assessment model described by Martin and Cloke, (2000) useful when considering factors that may affect consistency when assessing trainee teachers on school placement.

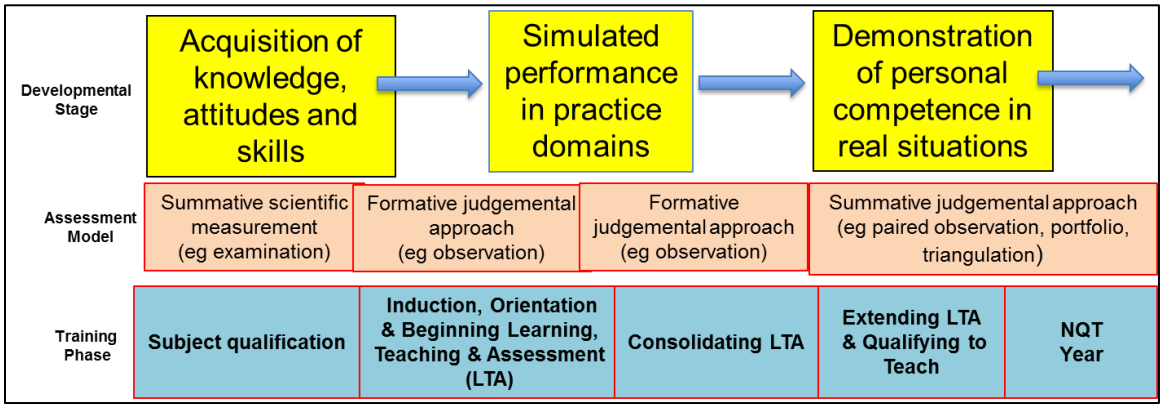

Figure 3. A model for the professional development of trainee teachers (developed from Hager & Butler (1996) and Martin & Cloke (2000).

The basic pre-requisite for aspiring teachers is adequate subject knowledge and trainee teachers in England must possess an honours degree in an appropriate subject or an equivalent qualification. Arriving at the start of teacher training, success in a subject at undergraduate level will have been judged using mainly scientific measurement assessment tools such as written assignments, portfolios and examinations (Hager and Butler, 1996). The current compulsory skill tests in English and Mathematics are further examples of filters applied to applicants for ITT/E programmes using a scientific measurement assessment tool.

Trainees extend their knowledge and skills beyond their own subject discipline as they progress through the various stages of their professional development (Figure 3). Shulman (1986) referred to subject knowledge for teachers in terms of three areas: subject matter content knowledge, pedagogical content knowledge and curriculum knowledge. Banks, Moon and Leach (2005) considered professional teacher knowledge to be a personal construct of subject and pedagogical knowledge together with school knowledge. The Teachers' Standards (Department for Education, 2011) include all these elements within the various standard descriptors that are used to assess trainee competencies. However, the assessment of trainees' knowledge on school placement is likely to involve qualitative judgements rather than formal testing (Martin & Cloke, 2000). As trainees practice and develop their teaching skills and then demonstrate competency (Figure 3) they take on increasingly independent responsibility for their classes.

Successful application of professional knowledge in the classroom becomes the assessment focus. Assessors then rely entirely upon a judgemental approach based upon qualitative evidence that is often considered less reliable and more subjective than scientific measurement (Martin & Cloke, 2000).

Leshem & Bar-hama (2008) investigated and discussed the issues that arose when criteria based assessment of teaching competencies was adopted by their ITT/E programme. Tutors used the criteria analytically or to confirm their overall holistic assessment decisions about teaching and learning. Their students perceived a role for clear assessment criteria and criterion based assessments during feedback but preferred holistic approaches to summative assessment.

The mentors and tutors in our study were expected to make evidence based judgements based upon qualitative evidence. However, holistic, analytical or combined approaches to assessment were all consistent with the framework and guidelines agreed with partner schools. Our interventions constituted new or amended organisational steps intended to reduce the potential for inconsistency between assessors due to subjective interpretation of assessment criteria and personal differences when applying the guidelines.

The assessment data for 2014-15 yielded evidence of a high degree of consistency across the five ITT/E programmes and their partnerships especially for the final summative assessment of the overall teaching grade. No firm conclusions can be reached at this stage about the reasons for this. It is, however, reasonable to speculate on the list of interventions and identify those which are associated solely with the final summative assessment. The interventions designed to counter inconsistency through assessor subjectivity can be summarised as:

- increased emphasis on mentor training,
- the application of rigorous, common assessment procedures based upon the Teachers' Standards (Department for Education, 2011) and grade descriptors developed and set down in the individual trainee standards tracking document, and
- the formalisation of final triangulation meeting procedures for quality assuring the summative assessment of trainees.

The evidence is circumstantial and causal links have yet to be established but the nature of the revised final assessment triangulation meeting with the presence of an external quality assurer are possible influences on the high degree of consistency of final assessment outcomes.

With this in mind the concourse of statements about the assessment of trainees on placement required for Q analysis (Brown, 1980, van Exel & de Graaf, 2005) was constructed. Just over forty statements based upon the Teachers' Standards (Department for Education, 2011), partnership documentation and training materials were selected with reference to the results of the quantitative phase. Mentors and tutors have been invited to carry out an on-line exercise to place the statements into a personal priority order. Q analysis of the results will identify clusters of subjectivity due to groups of respondents with differing assessment priorities. It is possible that this may provide a link to one of the interventions put in place to encourage consistency in outcomes and practice or may identify a different influence.

Conclusions and recommendations:

All the statistical methods trialled indicated consistency of overall teaching grade assessment outcomes across all programmes for the final summative assessment just prior to the recommendation of QTS. This constitutes strong evidence of consistency in assessment outcomes across all programmes. Future comparisons should be routinely performed at each formal review point as part of quality assurance procedures. Chi squared analysis based upon assessment data and contingency table calculations is the method recommended from those trialled in this study.

The comparison of the grades awarded for individual teachers' standards and the overall teaching grade using correlation coefficients demonstrated only strong positive correlations. This gives some indication that assessors are keeping to the guidelines provided. As there were no differences in the conclusions reached using parametric and non-parametric methods, applying the quicker of the two methods, Pearson's Correlation Coefficient is recommended after each review point.

The rigorous and formal nature of the triangulation meeting that confirms final assessment judgements should be retained until there is evidence to the contrary that this has contributed to the consistency of assessment outcomes at this point.

References

Brown, S. R. (1980), *Political Subjectivity: Applications of Q Methodology in Political Science. Yale:* Yale University.

Banks, F., Leach, J. & Moon, B. (2005) 'Extract from new understandings of teachers' pedagogic knowledge', *The Curriculum Journal*, 16(3), pp.331-340.

Burton, D. & Bartlett, S. (2009) *Key Issues for Education Researchers*. London: Sage.

Department for Education (2011) *Teachers' Standards: Guidance for School Leaders, School Staff and Governing Bodies*, Crown copyright 2013 Available at: www.gov.uk/government/publications/teachers-standards (Accessed 23 November 2016).

Van Exel, N. Job A. & de Graaf, G. (2005). Q methodology: A sneak preview. Available at: www.jobvanexel.nl] or visit http://www.qmethodology.net (Accessed 23 November 2016).

Hager, P. and Butler, J. (1996) 'Two models of educational assessment', *Assessment and Evaluation in Higher Education*, 21(4), pp. 367–378.

Leshem, S. & Bar-Hama, R. (2008) 'Evaluating teaching practice', ELTJournal, 62(3), pp.257-265.

Martin, S. and Cloke, C. (2000)' Standards for the Award of Qualified Teacher Status: reflections on assessment implications', *Assessment and Evaluation in Higher Education*, 25(2), pp. 183–190.

Norman, G. (2010) 'Likert scales, levels of measurement and the ''laws''of statistics', *Advances in Health Science Education*, 15, pp.625–632.

O'Donovan, B., Price, M. and Rust, C. (2001) 'The Student Experience of Criterion-Referenced Assessment (Through the Introduction of a Common Criteria Assessment Grid)', *Innovations in Education and Teaching International*, 38(1), pp. 74-85.

Ofsted (2015) *Initial teacher education inspection handbook*. London: Crown.

Philpott, C. (2014), T*heories of Professional Learning: A critical guide for teacher educators.* Critical Publishing.

Shulman, L. S. (1986) 'Those who understand: knowledge growth in teaching', *Educational Researcher*, 15(2), pp.4-14.