Can effect sizes give any clue to the way mentors ascribe numerical grades when assessing trainee teachers against the teachers' standards in England?

Teacher Education Advancement Network Journal

Vol 11(1) pages 4-14

Rick Tynan and Robert Bryn Jones
Liverpool John Moores University

Abstract

Some teacher educators use numerical grades when assessing teaching competencies. In this situation, statistical analysis can be used to monitor consistency and look for correlations between assessment outcomes across teacher training partnerships and at different stages in training. Another approach is to calculate effect size metrics. These do not claim statistical significance but do seek to explain the practical impact of patterns in quantitative data. This study looks at number grade assessment data from a large secondary initial teacher education programme across schools working in partnership with a higher education provider in the Northwest of England. The proportion of variance between numerical grades for individual Teachers' Standards and overall teaching was calculated at each formal review point over three consecutive years. Despite the complex process involved in assessing teaching competencies against performance criteria and the potential for subjective variation between individual assessors, the data consistently demonstrated underlying patterns. These suggested that quality assurance and management of assessment issues could have been a major influence on the assessors.

Key words

ITE; assessment; competencies; grades; effect size; secondary; standards; criteria; mentors; partnership.

Context and Review of Literature

Currently, Initial Teacher Education (ITE) programmes in England take place largely or entirely in schools, academies and colleges in partnership with providers who can accredit Qualified Teacher Status (QTS). This study is located in partnerships between secondary schools in the Northwest of England and a single Higher Education (HE) provider. School based mentors had first responsibility for both training and assessing trainee teachers subject to moderation and quality assurance by the provider. In England, HE and other providers are responsible for accrediting recommendations for the award of QTS. Such recommendations are based upon trainees demonstrating the minimum performance criteria described in the eight areas of teacher competency and section on professional expectations that are set out in the Teachers' Standards (Department for Education, 2011). The purpose of this investigation was look for clues in quantitative grading data to the priorities assessors gave to individual standards when considering overall teaching grades.

ITE partnerships in England are inspected and monitored by a government agency, the Office for Standards in Education (Ofsted). Ofsted judges ITE providers according to trainee outcomes (retention, grades and employment rates), the consistency of their training experience acrosspartnerships and the accuracy of mentors' assessments (Ofsted, 2018). The Northwest of

Citation

Tynan, R. and Jones, R.B. (2019) 'Can effect sizes give any clue to the way mentors ascribe numerical grades when assessing secondary trainee teachers against the teachers' standards in England?' *TEAN journal,* 11(1), pp. 4-14.

England HE provider in this study used Ofsted number grades for both formative and summative reviews of trainees' teaching skills in order to monitor and demonstrate their progress. Assessors all followed one possible assessment practice in England by numerically grading the eight individual standards and overall teaching at several formal review points during training. They used a four-point scale: 1 (Outstanding), 2 (Good), 3 (Requires improvement) and 4 (Inadequate). Although not all ITE providers use number grades to formatively assess their trainee teachers' performance, in England they must provide Ofsted inspectors with summative assessments of their trainees' teaching performance. In turn they are, themselves, judged on their ability to produce Good (2) and Outstanding (1) teachers (Ofsted, 2018).

In an effort to improve the consistency of assessment practice across a large number of partnerships, the HE provider in this study adopted some changes in 2011. These intended to improve the quality assurance of assessment practices. The provider also sought ways of demonstrating and monitoring consistency in the assessment grading outcomes of trainees. The steps taken were:

- Increased participation in mentor training by delivering this in partnership schools in addition to centrally, at the provider.
- Insisting on the central role of an agreed set of performance criteria, contained in an individual trainee standards tracking document, when making grading judgements against the Teachers' Standards (Department for Education, 2011).
- Adopting a rigorous and structured format for triangulation meetings between the trainee and mentor. These were chaired by a tutor from the provider and considered the evidence for the trainees' final indicative grades.
- Training tutors from the provider to emphasise their quality assurance and mentor training roles when visiting partnership schools.
- Agreeing clear documentation and guidelines through partnership steering groups.
- Emphasising preparation for inspection during mentor and visiting tutor training using feedback from external examiners and an Ofsted consultant.
- Using statistical tests to monitor consistency in grading outcomes and using this to inform training.

Practitioner researchers at the provider have found quantitative evidence of consistency that masks subjectivity in graded assessment outcomes across the partnerships monitored. Tynan and Mallaburn (2017) assumed that consistency in assessment practice would be reflected by consistency in assessment outcomes. They explored the use of statistical tests of significance to monitor consistency in numerical grades awarded by school based assessors. They demonstrated significant positive correlations between grades awarded for individual standards and the grades awarded for overall teaching. They also found consistency in final summative grades awarded for overall teaching across five ITE programmes (Tynan and Mallaburn, 2017). Whilst accepting there could be other explanations, Tynan and Mallaburn (2017) attributed their findings, at least in part, to the package of the interventions described above that were introduced to improve consistency of practice between assessors.

Tynan and Jones (2018) were interested in the assessment of trainees' subject knowledge on a secondary ITE programme and chose to focus on grades awarded for standards that include different aspects of subject knowledge for teachers. They used the most sensitive test of statistical significance

indicated by Tynan and Mallaburn's (2017) study to look at the relationship between grades awarded for Teachers' Standards S3 and S4 (Department for Education, 2011) and overall teaching grades in English, mathematics and science. Again, there was much consistency in the core subjects but in science and mathematics there were occasions when significantly more high grades were assigned for the standard associated with subject content and curriculum knowledge compared to overall teaching or the standard more associated with pedagogy (Tynan and Jones, 2018). This hinted at some subjectivity between assessors in different core subject areas not readily apparent in Tynan and Mallaburn's (2017) wider survey and trial of statistical analyses.

However, there are a number of issues associated with the use of Ofsted number grades when assessing trainee teachers' performance that make achieving consistency and accuracy across a large number of partnerships problematic. The Teachers' Standards (Department for Education, 2011) give information on the minimum performance criteria necessary for the recommendation of QTS in England. However, the Teachers' Standards (Department for Education, 2011) contain neither guidance on appropriate assessment tools nor acceptable evidence to be used for assessments. Further, they do not contain any information on criteria for judging Good (2) or Outstanding (1) performance. ITE providers must use locally agreed criteria to evaluate performance above the minimum required. At the provider in this study, these were formulated initially by a consortium of local bodies involved in ITE. Over time, the provider has developed these with partnership schools using extrapolations from the Standards descriptors, Ofsted descriptions of the characteristics of undergraduate final year trainees and, more recently, clues from Ofsted ITE partnership inspections. Inherent in this approach is the opportunity for subjective differences between regions and local partnerships in the choice of assessment tools, construction and interpretation of criteria and choice of evidence when assessing trainee teachers.

In addition to the potential sources of variability inherent in basing assessments on Ofsted categories coded as numbers, there are some theoretical issues that predict that more variability in assessment grades might be expected than the practitioner investigations cited above actually demonstrated. The way in which assessors perceive professional learning could have important implications for their approach to assessment and pose another potential source of subjectivity and variation in grades. Philpott (2014) provided a summary and critical review of a range of models for professional learning and their implications for teacher educators. These constituted a continuum with individual cognitive and psychological approaches at the opposite end to those that consider learning to be a social construct. For instance, Kolb's (1983) model, which focuses on the acquisition of knowledge and skills through experiential learning, seems to invite the assessor to concentrate on the aspiring teacher's performance. On the other hand, Wenger's (1998) model, which emphasises acceptance into a community of practice, suggests judgements based upon norms, expectations and aspiring practitioners' perceived impact on learners as clients (Philpott, 2014).

Hager and Butler (1996) considered two models of assessment to be necessary during professional learning and Martin and Cloke (2000) applied these to the assessment of teaching competencies. They contended that, whatever model of professional learning is assumed, the evaluation of trainee teachers becomes more judgemental and qualitative and less measurable scientifically as professional learning proceeds. Tynan *et al* (2014) compared the assessment outcomes of Post Graduate Certificate in Education (PGCE) science trainees preparing to teach chemistry and physics. Some arrived at the HE QTS provider with a first degree in their specialist teaching subject whilst others arrived with a one year Subject Knowledge Enhancement (SKE) qualification accredited by an HE Certificate. Tynan *et al*

(2014) found that school based assessors did not distinguish between these two groups of science trainees and that the grades awarded for subject knowledge and overall teaching ability were similar no matter the level of qualification in chemistry or physics. As it would seem impossible to cover as much science subject content in one year compared to a three year undergraduate course, these findings would seem to suggest that assessors were assessing subject knowledge in a different manner to that used at the end of a first degree. This would seem to validate the application of Martin and Cloke's (2000) model in that context. Tynan and Mallaburn (2017) mapped this model to the delivery and assessment of ITE programmes at one HE provider and noted the implications for increased variability in grades if the model was assumed to be valid.

Criteria based assessment of competencies can be viewed as an approach aimed at reducing the subjectivity inherent in a qualitative judgemental assessment model. Leshem and Bar-Hama (2008) discussed issues around the introduction of criteria based assessment of teaching competencies to the Israeli ITE system. They explored the attitudes, perceptions and preferences of tutors and students in comparing this more analytical approach to the previous practice of using professional judgement more holistically. Tutors in their study reported different approaches to using criteria for assessing teaching competencies. Some started with a holistic judgement and used the criteria as a check, whilst others started more analytically with the criteria and then compared the resulting assessment outcome against their professional judgement. No matter their approach, tutors noted difficulties in reconciling their holistic judgments with criteria based assessment. In the context of this study, assessors following the guidelines agreed by partnership schools and the HE QTS provider graded individual standards first and then used these to arrive at an overall grade for teaching. However, as Leshem and Bar-Hama (2008) reported, some assessors may find this approach difficult and could have started with a holistic assessment and grading of overall teaching ability and then graded the individual standards accordingly afterwards.

Tummons (2010a, 2010b and 2011) has also considered in depth assessment across PGCE programmes provided by a northern university. Tummons (2010b) investigated the validity and reliability of assessments of trainee lesson plans and also the issues associated with making valid assessments of trainees' reflective practice (Tummons, 2011). However, when considering possible reasons for consistency in number grade assessments, Tummon's (2010a) application of institutional ethnography (IE) and actor network theory (ANT) to the assessment of post graduate trainees appear useful. This approach perceives assessment as closely governed by IE and ANT. IE can be described as the way an institution documents its courses and assessment activities which, in turn, becomes inseparable from the way these documents are sponsored by tutors and teachers (ANT). The application of this approach to student teacher assessment led Tummons (2010a) to suggest that complex assessment activities had been subsumed in practice by quality assurance and managerial issues. These ideas might help explain the high levels of consistency in grading assessment data reported by Tynan and Mallaburn (2017) and Tynan and Jones (2018) for competency based assessment against the Teachers' Standards (Department for Education, 2011).

Methodology and methods

This study constitutes local, small scale, practitioner research involving one secondary ITE programme at a single HE QTS provider in the Northwest of England. It is a quantitative survey and analysis of numerical grades for individual teaching standards and overall teaching. These were collated from formal progress review forms during the period September 2014 to July 2017. The programme selected was the largest of those available for study at the QTS provider, which earlier work (Tynan

and Mallaburn, 2017) suggested was representative of the other programmes. School based mentors routinely assessed trainees preparing to teach a range of subjects during their Post Graduate Diploma in Education/Certificate in Education (PGDE/CE) courses. All the trainees were in secondary schools. Trainees were assessed against the Teachers' Standards (Department for Education, 2011) using the descriptors in the standards and a locally produced trainee progress tracking document used across all the partnerships within the programme. Numerical grading data on formal review forms were collected from three consecutive cohorts.

Quantitative studies based upon tests of statistical significance can be criticised if they omit to attempt an explanation of the practical significance, or impact in everyday terms, of their statistical findings (Ellis, 2010). Proportion of variance (POV) is one approach to addressing this using an r-family effect size metric that looks at the practical impact of correlations (Ellis, 2010). A further advantage of using an effect size metric is that the statistic is scale free. This allows comparison of data from different studies (Ellis, 2010) and, in this study, the data collected from three different years, despite differences in cohort sizes.

Previous work by Tynan and Mallaburn (2017) on numerical grades established that the use of either Pearson's or Spearman's Rank correlation coefficients (r) led to identical statistical conclusions (for a starter statistical text see Hinton, 2014). No matter which test was used, positive correlations were demonstrated between the grades awarded for individual standards and overall teaching, with very low probabilities of these being due to random patterns in the data (Tynan and Mallaburn, 2017). In light of this finding, the most straightforward calculation was adopted. Pearson's correlation coefficient (r) was calculated using the standard function formula available in standard spreadsheet software. Grades for each individual standard were compared to grades for overall teaching at every formal review point where number grades were ascribed. POV (r^2) was calculated by squaring r (Ellis, 2010) using a standard spreadsheet formula. This effect size metric can be reported as the proportion of variance or as expressed as the percentage of the variation shared by two sets of data, simply by multiplying r^2 by one hundred (Ellis, 2010). Percentages may be more intuitively understood than proportions written as decimals.

Considering POV does not seek to prove statistical significance but to establish a practical indication of the size of an effect, no matter the cause. The aim of this study was to use POV to establish if the grades assigned for some individual standards might be linked more closely than others to overall teaching grades. In turn, this might give clues to the priorities ascribed to different standards by assessors when deciding on an overall grade for teaching.

Several qualifications should be noted before considering the findings below. POV metrics are calculated from correlation coefficients. However, demonstrating correlations between sets of grades does not describe causal reasons for them, if any exist. Further, even if the correlations underlying the calculations of POV are significant, any discussion of the differences between effect sizes must include the distinct possibility that these might be the result of chance variation in the data. Lastly, the cohorts were treated as full populations not samples, so there is a high degree of validity in the data for the programme investigated but no claim that the findings from this programme should be extrapolated to a larger population of trainee teachers.

Findings

Table 1 provides a quick reference to the Teachers' Standards headings (Department for Education, 2011). In the interests of reducing the amount of statistical data presented, only the results for the

final summative assessment point are presented in full (Table 2 and Figure 1 below). However, in Table 3 (below), for every review point, the pairs of standards whose grades shared the highest and lowest percentage variation overlap with the grades for overall teaching are presented. The difference in the percentage overlap from first to last ranked standard is also reported as an indication of the range of the effect sizes.

Table 1: Key to Part 1 Teachers' Standards headings (Department for Education, 2011)

1. Set high expectations which inspire, motivate and challenge pupils.
2. Promote good progress and outcomes by pupils
3. Demonstrate good subject and curriculum knowledge
4. Plan and teach well-structured lessons
5. Adapt teaching to respond to the strengths and needs of all pupils
6. Make accurate and productive use of assessment
7. Manage behaviour effectively to ensure a good and safe learning environment
8. Fulfil wider professional responsibilities

Main Findings

There were differences each year between the POV for summative grades ascribed for individual standards and overall teaching (Table 2 and Figure 1)

Table 2: The percentage proportion of variance (POV) for the final summative grades of individual standards compared to grades awarded for teaching overall and their yearly rankings.

Teachers'	Proportion of variance (%)			Ranking		
Standard	2014-15	2015-16	2016-17	2014-15	2015-16	2016-17
1	65	56	64	3	4	3
2	70	60	68	1	2	2
3	27	24	47	8	8	6
4	66	64	69	2	1	1
5	52	57	57	5	3	4
6	48	52	51	6	5	5
7	54	50	43	4	6	7
8	41	35	37	7	7	8
n	85	126	238			

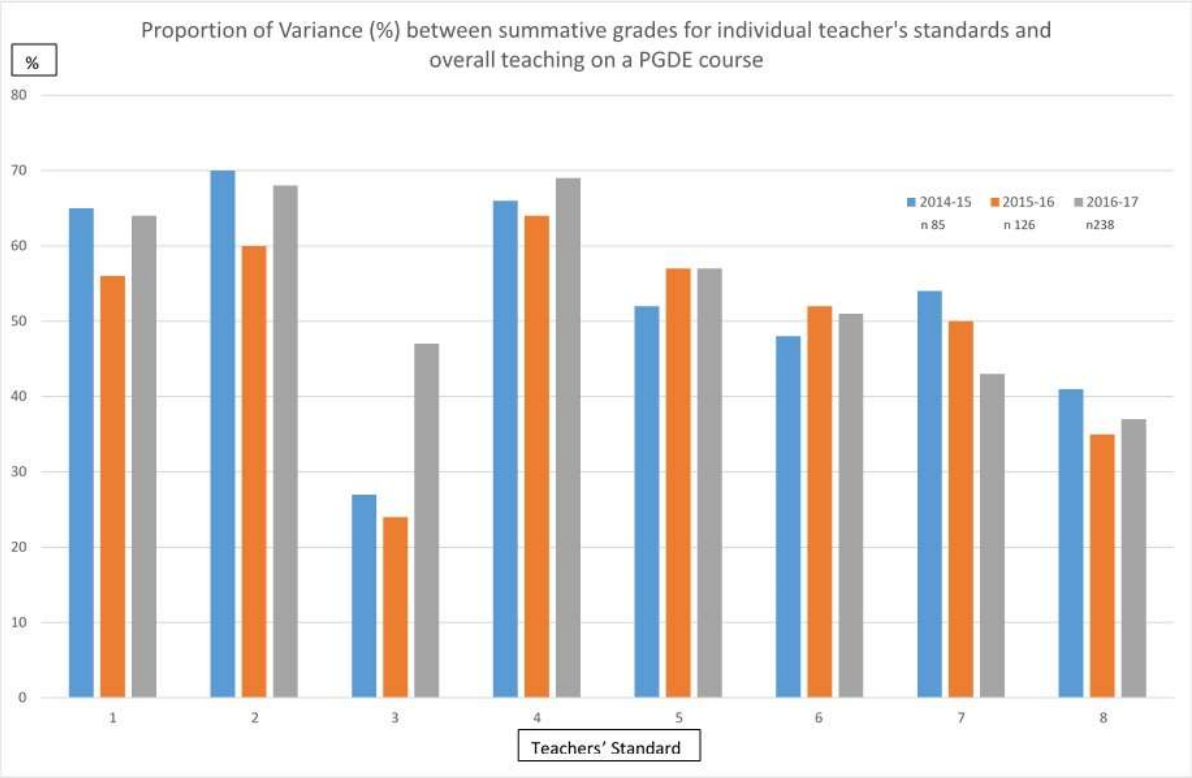

Figure 1. The percentage of shared variation between summative numerical grades ascribed for individual Teachers' Standards (Department for Education, 2011) and overall teaching over three consecutive years.

At all review points where number grades were ascribed, there were patterns in the POV for grades for individual standards and overall teaching that were similar over a three year period (Table 3).

Table 3: Pairs of Teachers' Standards (Department for Education, 20011) with the highest and lowest proportion of variance (POV) for a Secondary ITE Programme leading to QTS for all review periods assessed with number grades.

	2014-15	2015-16	2016-17
Highest Ranking Standards			
First formative review	1 & 4	1 & 5	2 & 4
Second formative review	2 & 4	-	-
Third formative review	2 & 4	1 &4	2 & 3
Final summative review	2 & 4	2 &4	2 & 4
Lowest ranking standards			
First formative review	6 & 8	6 & 8	6 & 8
Second formative review	3 & 8	-	-
Third formative review	3 & 8	3 & 6	7 & 8
Final summative review	3 & 8	3 & 8	7 & 8

For formative review points during 2014-2017, Standards 1, 2 and 4 were most likely to share the highest percentage of their variation in grades, with overall teaching and standards 3, 6 and 8 most likely to share the least (Table 3). For final summative grades during 2014-2017, Standards 2 and 4 always shared most variation with overall teaching and Standards 3 and 8 were most likely to share the least (Table 3 and Figure 1).

One possible interpretation of these findings is that the different effect sizes are caused by non-random differences in the way that number grades were ascribed. If this were the case then the differences in effect sizes could indicate differences between the ways assessors perceive different standards' contribution to overall teaching performance. For instance, for summative grades, one interpretation could then be that assessors associated the grades for Standard 2 and Standard 4 much more closely with the grade ascribed for overall teaching than the grades for Standard 3 and Standard 8 (Table 1). This approach does not help identify reasons why this might be the case, but the consistency of this pattern over three years encourages the idea that the pattern is non-random and a further investigation into the possible reasons for the pattern would be worthwhile.

Discussion

The interpretation of POVs as percentages (Ellis, 2010) allows an easily accessible interpretation of the practical significance of the correlation between the grades for a particular standard and overall teaching. For example, in the first line of Table 2 the POV for Standard 1 was found to be 64% in 2016-2017. This can be interpreted as 64% of the variation found in the grades for overall teaching and Standard 1 was common to both. Intuitively, this would seem to indicate that assessors were placing more importance on Standard 1 than Standard 8 which only shared 37% of its variation with overall teaching. Of course, this is a risky interpretation as neither correlations nor shared variation can be used alone to establish causal relationships. Further, the use of the effect size metric invites the interpretation that 64% of the variation in overall teaching grades was due to variation in the grades for Standard 1, whereas only 37% was due to variation in grades for Standard 8. This is one possible explanation but assumes that the differences in POV were due to non-random causes and that assessors had followed the agreed guidelines by using a profile of grades for individual standards to arrive at an overall teaching grade. Neither may be the case.

However, it is not necessary to make any conclusions about potential causal relationships for correlations for their effect size metrics to be useful tools. When tracking consistency in assessment outcomes and practices it is sufficient to know that there may be a trend that needs further monitoring whilst more qualitative evidence is gathered. Effect sizes are scale free and can be compared directly across different data sets collected during an investigation (Ellis, 2010) or the meta-analysis of data from different studies (Cooper, 2017). This allows the comparison of the effect sizes for different standards in the same year and also effect sizes for the same standard in different years. For example, in Table 2, in the academic year 2014-15, POV values suggest that assessors linked Standard 4 most often and Standard 8 least often with overall teaching grades. Using Table 1, this might suggest that assessors mentally associated a trainee's pedagogical knowledge more often with overall teaching ability than their contribution to wider school responsibilities. Also from Table 2, it can be seen that the POV values for Standard 1 differ in different academic years. This may indicate changing attitudes to the relative importance of this standard and other standards over time. However, such ideas have to be considered very cautiously as variations in effect size metrics could still represent chance fluctuations in data (Ellis, 2010).

The consideration of holistic and analytical competency based assessment practices by Leshem and Bar-Hama (2008) and Hager and Butler's (1996) discussion of the qualitative judgemental assessment model, applied to education by Martin and Cloke (2000), both appear to indicate that a degree of subjectivity amongst a large group of assessors should be expected. Similarly, differences in assessors' views on the psychological (Kolb, 1983) or social (Wenger, 1998) nature of professional learning could also lead to subjective differences in the way they ascribed numerical grades. A consideration of POVs has indicated more variation than was demonstrated by the use of statistical tests of significance by Tynan and Mallaburn (2017) and Tynan and Jones (2018). The variation in POV values between standards and the same standards over time may suggest the subtle effects of differences between assessors on grading outcomes. However, there were also consistent patterns in the POVs calculated and one possible explanation may lie in the interventions listed previously that were successfully applied during the period of this study. These encouraged a standards-first approach and sought to improve consistency of assessment practice across the school and HE provider partnerships involved. However, it is interesting to note that the steps implemented focused upon agreed procedures and consistency of outcomes. Assessors and liaison tutors were not formally called upon to consider or question the assessment process by which judgements and number grades were assigned.

A consideration of the contents of Table 3 suggests that this pragmatic approach may have affected the way assessors ascribed grades for individual standards and overall teaching. During the period 2014-2017 the partnerships were awaiting inspection by Ofsted and increasing emphasis was placed on the interventions described previously that were implemented to improve the quality of partnerships in line with published Ofsted ITE inspection criteria (Ofsted, 2018). Participation in training for mentors of trainees in school greatly increased and the reference point for awarding grades was the descriptors in the Teachers' Standards in England (Department for Education, 2011) and a locally agreed set of performance criteria. It was considered very important that summative grades were agreed at rigorous and structured triangulation meetings involving trainees, their mentors and the visiting tutor from the HE QTS provider in a quality assurance role. There can be little doubt that information considered important to establish consistency in preparation for inspection was cascaded repeatedly to mentors.

The three year patterns in POVs for summative grades are congruent with the information disseminated. Namely that it would be difficult to justify numerical grades for overall teaching that were widely different to the grades of some important standards. Standards 2 and 4 were considered the best indicators of the overall teaching grade. Standards 5 and 6 were also considered to be important predictors. This is not to suggest that this is or ever was actual Ofsted practice during inspections of ITE partnerships in England but merely to record the advice that was cascaded to all interested participants during the period of this study.

Tummons (2010a) found it useful to consider IE and ANT when discussing assessments in ITE. Tummons did not look specifically at competency based assessment of trainees on teaching experience placement by school mentors against the Teachers' Standards (Department for Education, 2011). However, the findings in Table 3 are congruent with the idea that the ITE programme's documentation together with its sponsors, the school liaison tutors and school mentors, were successful in sharing and implementing the programme's messages on consistency in grading in preparation for Ofsted. For formative review point grades, other standards might be included in the top pair and an inspection indicator standard might be included in the bottom pair. However, for every

year during the study, the summative grades for Standards 2 and 4 shared the highest POV with overall teaching and the summative grades for Standards 5 and 6 were in the middle group and never in the lowest pair.

It is difficult to imagine that chance fluctuations might give rise to the same pattern over three consecutive years. This might indicate that, whilst some assessor subjectivity was possible during formative grading, for summative assessments quality assurance and assessment managerial issues could mask the assessment process and its associated sources of variation. This could constitute a further example of IE and ANT in ITE assessment similar to that suggested by Tummons (2010a).

Conclusions

The use of an effect size metric for the quantitative investigation of grades assigned for individual standards and overall teaching has suggested some findings about the way assessors in schools ascribe grades and raised further questions. These are highly relevant to practitioners in the programme and institution studied and may be of general utility to all teacher educators involved in assessing trainee teachers against criteria describing teaching competencies.

In this study, the number grades assigned did not reflect the degree of variability predicted by several procedural and theoretical considerations. Further, consistency in numerical grading outcomes may have reflected consistency in assessment practices between assessors but there is a suggestion that compliance with quality assurance and management of assessment issues also contributed. In the minds of the authors, this begins to build a case questioning the use of numerical grades for individual standards and overall teaching during the formative and summative assessment of trainee teachers.

Professional learning deserves a valid and reliable assessment process with the aim of assessment to produce the most effective teachers possible. Given the issues discussed previously concerning the application and extrapolation of the Teachers' Standards (Department for Education, 2011), ascribing number grades or categorising teachers' teaching performance in these circumstances invites the subjective use of professional judgement whilst leading an external observer to believe that this is something that can be measured scientifically (Martin and Cloke, 2000). The Teachers' Standards (Department for Education, 2011) may provide a useful analytical approach for mentoring and coaching aspiring teachers. However, the use of number grades may not facilitate meaningful assessment and, particularly in the formative stages of an ITE programme, may reduce the validity and utility of feedback to trainee teachers.

At the HE provider studied, statistical analysis has been used to demonstrate high levels of consistency in numerical grading outcomes across partnerships, programmes and time (Tynan and Mallaburn, 2017, Tynan and Jones 2018). The findings of the current study give clues that this may be due to the effects of successful managerial and quality assurance interventions that ensure compliance with assessment guidelines rather than reducing subjective differences in assessment practice between assessors. In their present format the use of number grades and categories when assessing teachers may be masking actual assessment processes, which may be more valid and reliable than the current practice.

Next steps

Further qualitative projects are in progress or in the planning stages at the ITT/E HE QTS provider that seek to answer questions that cannot be addressed by further quantitative studies:

- How far can the standards descriptors be trusted to guide assessment of trainees when considering performance above the minimum required for QTS?
- What are the tensions perceived between assessment and its practice?
- Is potential subjectivity between assessors an issue?
- How do assessors perceive assessment using performance criteria?
- What evidence do assessors use to ascribe grades and how do they use it?
- Does a number grade approach lead to spurious perceptions of accuracy in, what is essentially, a qualitative assessment system?
- Can quality assurance and management considerations across many partnerships allow complex assessment processes to be implemented fairly and reliably?

References

Cooper, H. (2017) *Research Synthesis and Meta-analysis: A step-by-step approach: 5th Edition*. 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

Ellis, P. D. (2010) *The essential guide to effect sizes: Statistical Power, Meta-Analysis, and the interpretation of Research Results*, Cambridge University Press.

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

Hinton, P. R. (2014) *Statistics Explained*, *Third Edition*, London: Routledge.

Kolb, D. A. (1983*) Experiential Learning: Experience as the Source of Learning and Development*, London: Prentice Hall.

Leshem, S. and Bar-Hama, R. (2008) 'Evaluating teaching practice', *ELT Journal*, 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), 183–190.

Ofsted (2018) *Initial teacher education inspection handbook*, Crown copyright

Philpott, C. (2014) *Theories of Professional Learning: A Critical Guide for Teacher Educators*, Northwich: Critical Publishing.

Tummons, J. (2010a) 'Institutional ethnography and actor network theory: a framework for researching the assessment of trainee teachers', *Ethnography and Education*, 5 (3), pp. 345-357.

Tummons, J. (2010b) 'The assessment of lesson plans in teacher education: a case study in assessment validity and reliability', *Assessment & Evaluation in Higher Education*, 35(7), pp. 847-857.

Tummons, J. (2011) 'It sort of feels uncomfortable': problematizing the assessment of reflective practice', *Studies in Higher Education*, 36(4), pp. 471-483.

Tynan, R. and Jones, R. B. (2018) 'Assessing trainee secondary teachers on school placement: subject knowledge and overall teaching grades', *TEAN Journal*, 10(1), pp. 20-34.

Tynan, R. and Mallaburn, A. (2017) 'Consistency Counts- or does it?' *TEAN Journal*, 9(1), pp. 90-99.

Tynan, R., Mallaburn, A., Jones, R. B. and Clays, K. (2014) 'Subject knowledge enhancement (SKE) courses for creating new chemistry and physics teachers: do they work?', *School Science Review*, 95(353), pp. 85–94.

Wenger, E. (1998), *Communities of Practice: Learning, Meaning and Identity*, Cambridge University Press.