Title Page

Title

A comparison of teacher assessment with standardised tests in primary Literacy and Numeracy: how Assessing Pupil Progress (APP) is used in schools and how it compares with the Wechsler Individual Achievement Test (WIAT-11).

Ruth Marlow[a]*, Brahm Norwich[b], Obioha Ukoumunne[a], Lorraine Hansford[a], Siobhan Sharkey[c] and Tamsin Ford[a]

[a]Child Mental Health, University of Exeter Medical School, Exeter, UK; [b]Graduate School of Education, College of Social Sciences and International Studies, University of Exeter, Exeter, UK; c Centre for Clinical Trials & Health Research – Translational & Stratified Medicine, Plymouth University, Plymouth, UK

(Received 25 November 2013; accepted 15 June 2014)

Abstract

Background: Assessing Pupils Progress (APP) arose from a government drive to increase the amount of teacher – based assessment within school and to make this consistent across schools. Numeracy, literacy and science were targeted and a number of standards and criteria were developed to help teachers assess the attainment of their pupils, although only numeracy and literacy criteria are widely used. The criteria provided for the government enabled teachers to assign an attainment level to their pupils.

We conducted semi-structured interviews with Head teachers to gaininsight into how their schools applied APP and we compared the APP levels for English and Maths, provided by teachers across 11 schools for 72 pupils to a standardised assessment (Wechsler Individual Achievement Test-11; WIAT). The WIAT assessment is age-normed and standardised.

Results showed that there was a strong correlation between the APP and WIAT – 11 for literacy but not for numeracy. Head teacher interviews revealed that APP is used differently across schools and at times are used in a way inconsistent with government guidance. Therefore this once again raises the question of how teacher assessments are used and their purpose. Areas that should be considered are; how moderation is used in schools, what role teacher assessments have in educating pupils and what is good practice in relation to these assessments. Also important is consideration of the clarity around teacher assessment for teachers and schools. Clarity over the function of assessments is of vital importance as is ensuring that assessments are meaningful to teachers,

Comment [HL1]: Its probably just me but I'm not sure what this sentence means? Do we mean provided by the gov rather than for them?

Comment [RM2]: Add something about the implications.

pupils, families and schools, especially in light of the weight that can be placed on teacher assessments for all these groups.

Keywords: Comparison, Assessment, Standardised, Assessing Pupils Progress, Teacher

Main Text

Introduction

Teacher assessment of learning and attainment continues to be an important policy and practice issue in the UK and internationally (Harlen, 2004a). The introduction of the National Curriculum (NC) in England and Wales in 1988 also brought an emphasis on national assessment criteria. This meant an increased focus on the different qualities of teacher assessment and standardised tests (NC), for example highlighting the importance of reliability, validity, cost (which is provided by standardised tests) whilst considering the wider impact on teaching, learning and attainment (DES, 1987). Debate about assessment policy and practice persists despite an increasing reliance on teacher assessment in many countries, both for high-stakes summative and systematic formative assessment (Johnson, 2013). For example in relation to the General Certificate of Secondary Education within the UK, although this is less common now. It is common for summative assessment to involve standardised tests and tasks at the end of a phase of learning to provide an estimate of overall attainment. By contrast, formative assessment is about the provision of feedback to teaching staff and pupils about progress in order to focus their teaching and learning. Formative assessment produces a picture of learner strengths and future needs and requires the exercise of professional judgement by teachers. However, assessments that are designed for formative use can sometimes be aggregated and used for summative purposes, as suggested by the influential working group that advised on National Curriculum assessment (DES, 1987). Although the UK Government of the day did not adopt this

approach in the National Curriculum (NC) assessment arrangements, teacher assessment was still used alongside standardised tests for different summative reporting purposes. Standardised tests have been used for national and school evaluation purposes, while teacher assessment has been used for communication between teachers and with parents/pupils.

The Qualifications & Curriculum Authority (2009) launched the Assessing Pupils Progress (APP) initiative as part of the Assessment for Learning (AfL) Strategy. The UK government at the time invested £150 million over three years 2008-211 to help schools in England take a strategic approach to classroom assessment, with the aim of securing good practice, and APP was part of this. This initiative arose out of the Department for Children, Schools and Families project 'Making Good Progress' (DCSF, 2009). The Making Good Progress initiative involved termly monitoring of pupil progress using 'assessment for learning' methods, one to one tuition, use of single level tests, progression targets and premium payments. The APP initiative, therefore, had formative purpose as it was designed to support teachers' professional judgements about their pupils' progress.

APP was developed over five years, in literacy, numeracy and science and standardised to provide a common language for talking about pupil attainment. It was not a statutory requirement, but has provided a reference point for teachers in relation to national standards. It was meant to put learners at the centre of the assessment process by the provision of a detailed profile of attainment and progress. APP was designed to be evidence – based and to be used two or three times a year. It was planned to replace rather than be additional to existing assessment arrangements, and to be a complete, school-wide system of assessment based on daily teaching and learning evidence.

The APP materials consist of:

- A Teachers handbook; which gives the context for assessment and introduces the APP tool. (Qualifications and Curriculum authority, 2010 a&b).

- Standards files which aim to help schools reach consistent and reliable judgements about NC and provide exemplifications of national standards (Qualifications and Curriculum authority, 2009).
- Assessment guidelines which set out assessment criteria for each NC level (Qualifications and Curriculum authority, 2009).

The importance of the conditions under which APP is applied was recognised from its pilot (NUT, 2009). Teachers needed time to become familiar with it and this required that they share their experiences of using APP as part of staff working together to develop whole school assessment practices. In a later review of the APP system, Ofsted (2011) assessed the impact of APP in 14 secondary schools and 25 primary schools. They concluded that the system strengthened the assessment process in schools, increased teachers accountability and increased consistency and accuracy of assessment. However, it is unclear how these findings were established, as there is no clear presentation of the methodology used, and the results are not backed up by the presentation of data or either statistical or qualitative analysis.

Although Ofsted (2011) suggest that the system improved moderation in schools, there were no formal systems for the introduction of the APP materials or their implementation in schools or any empirical testing to support this assertion. The lack of formal processesmay have led to considerable variation between individuals using the system and the potential for considerable bias (Harlen, 2005). The academic literature indicates that in order to be able to rely on teacher assessment as part of a summative process, there would need to be a greater emphasis on training and ensuring consistency between teachers and between schools (Black et al., 2011).

There has been a long history of debate about how the process of teacher assessment may be subject to individual teacher differences. Wyatt-Smith & Klenowski (2013) highlight how many sources of information are used in creating judgements about a child's ability, including the teachers' individual knowledge and values, as well as socio-cultural factors. However, some would

argue that this makes teachers' judgements more valid than formal assessments due to their holistic nature (Allal, 2013). Others have said that teacher assessments are less vulnerable to variations within a child, such as a good or bad day on the day of formal assessment (Durant, 2003). In contrast, formal assessments can be perceived as more accurate and objective than teachers' assessments. Black et al. (2011) outline the historical distrust in teachers' assessment; some perceive that teachers may lack the sufficient knowledge and skills to complete assessments reliably.

The reliability of teacher assessments are important for schools, teachers, families and individual pupils; also because of the cost of external assessments in schools. Currently external assessments of children in primary school cost £24.31 million a year (Department for Education, 2013). If teachers' assessments were demonstrated to be accurate and reliable it may question the need for these external assessments.

There has been a paucity of research in this area. Durant (2003) compared the difference between Key stage assessment results (at that time key stage 1, 2 and 3 when all were externally assessed) and teacher assessments both expressed as NC levels, and found a differing level of consistency between teacher assessment and key stage tests, which were greatest at the key stage 2, ages 7-11, among 32,000 students over five years. They reported that 75% of teacher assessment and test levels at key stage 2 were identical, and a further 20% were within one test level. For Key stage 1 they found that 50% of them were identical and a further 35% were within one test level. At Key stage 3 they reported that 90% of teacher assessments/ test levels were either identical or within one test level. The percentage of identical agreement at this level was not reported. Although this study appears to indicate a relatively high level of agreement, it also suggests that agreement may vary with the age of the child. There was no control for whether teachers had used the external test results to inform their teacher assessments; in fact it was stated that teachers would have had to 'review' test scripts, which might have increased the level of agreement obtained.

Aim

To compare teachers' assessment of their primary aged pupils' attainments in literacy and numeracy using APP with their children's performance on the Wechsler Individual assessment test (WIAT-11). We also aimed to understand how APP was used as a tool within local schools.

Method

Ethical approval for the study was granted by the University of Exeter Medical School Research Ethics Committee. Figure 1 illustrates the participation of schools within the qualitative and quantitative aspects of this study.

1) WIAT- APP comparison

Although the APP criteria span Literacy, Numeracy and Science, consultations with local education professionals highlighted that Science APP criteria are seldom used in primary schools; as was also noted nationally by Ofsted (2011). Anecdotal evidence suggested that teachers found it too difficult to apply the criteria outlined for science. Therefore, only literacy and numeracy were examined.

We carefully studied and compared various academic assessment tools in order to choose a measure that best mapped the concepts used to generate teacher APP assessments in literacy and numeracy. The assessment tool needed to be; standardised and normed in the UK, measure broadly the same areas as the APP and be applicable to the age group assessed (4-9 years). Most tests were eliminated because they did not assess numeracy (*Table 1*).

Insert Table 1 here

The WIAT-II was found to most closely map onto the APP criteria (as outlined by the Qualifications curriculum authority, 2010 a & b; *Table 2*). The WIAT-II is a widely-used assessment of academic abilities, it provides normative data for ages 4 to 16 years 11 months, and was developed in the US and standardised in the UK with data from 892 individuals. The recruitment of the UK sample was stratified (from the UK census) to ensure representativeness from all demographic groups (based on

region, gender, age, race/ethnicity and parental education), which was then linked to the 1,069 participants sampled in the US. In development, the WIAT showed sound estimates of the standard error of measurement and high levels of reliability. Reliability scores indicate a strong inter-item consistency with coefficients ranging from .80 to .90. Data gathered also reports an adequate stability across age and time, using retest coefficients, these typically lie between 1 and 4 standard score points.

Insert Table 2 here

Unfortunately, no assessment reviewed would adequately assess the concepts related to speaking and listening that the APP criteria outline. Overall the fit between the criteria most closely match in the numeracy, although there was no element in the WIAT that assessed measurement and the WIAT had a greater focus on more complex problem solving. Part of the literacy assessment for the WIAT examined children'sability to read stories aloud and then to be able to use the knowledge gained from the stories to identify information from it. However, there is less focus in this assessment, when compared to the APP criteria, on the ability to infer and interpret from written text. Although spelling was assessed, the ability to write sentences was not.

The WIAT provides raw scores, percentiles that indicate what percentage of the general population would be expected to be performing above or below that level, and scaled scores which enable you to compare that child's performance to the performance that would be expected as compared to the normative data gathered. Using this it is possible to determine whether the child is performing at expectation (average), below/ above average, or significantly above/ below average.

Sample

This study is nested within a cluster randomised controlled trial of the Incredible Years Teacher Classroom Management course (Ford et al, 2012). The parent study, Supporting Teachers and childRen in Schools (STARS) recruited mainstream, state funded primary schools (children aged 4-

11). Data for this study was gathered in 2013. These schools were mostly located within an urbanised coastal area that experiences high levels of deprivation and social challenges. Schools on special measures were excluded as were schools that had no substantive head teacher. The study teacher had to teach a single year group class with at least 15 children and, be teaching the children for at least four days per week. The head teacher consented for the school to participate, and nominated a teacher who also consented. Parents were able to opt themselves and their child out of the main study. As part of this study parents were asked if they would be interested in their child participating in extra literacy and numeracy assessments; 243 parents out of 387 parents indicated that they were interested. The children of these parents were then divided into those functioning above/ below and at the expected level for their age group according to the teacher's allocated APP score. Forty children from each ability level (120 in total) were then selected at random. These parents were then approached for formal consent, of these 72 responded and were assessed (Table 3.) These children were from years 1-4 (age 5-9 years) and attended 11 primary schools in Devon. Children provided verbal assent on the day of assessment; distress or reluctance was interpreted as withdrawal of consent.

Procedure

The WIAT-II was administered in February 2013, in order for teachers to have time to get to know the child and provide a second APP assessment for the comparison at the time of the WIAT assessments.

The WIAT assessments were completed by five psychology undergraduate students who were trained and supervised by a clinical psychologist (RM). Results were fed back to the parents through a summary report that they were encouraged to share with teachers and anyone they felt would benefit from knowing the results (see appendix 1).

Data analysis

Characteristics of the participating children and teachers were summarised using means and standard deviations (or medians and interquartile ranges) for quantitative variables and percentages for categorical variables. The APP levels and sub-levels were converted into National Curriculum Points scores using conversion tables provided in Appendix A of Progression (DFE, 2010). Reported findings are based on these continuous point scores. The literacy and numeracy scores were summarised for each of the APP and WIAT measures. The strength of association between the APP and WIAT measures was quantified using the Pearson correlation coefficient for literacy and for numeracy. Tests of interaction using linear regression were used to investigate whether there was evidence that the association between the APP and WIAT measures differed between children taught by teachers who had taught for less than five years and those who have practiced for longer based on the timing of the introduction of the APP, and differed between children with and without special education needs.

2) Head Teacher Interviews

All the head teachers of the 15 schools recruited to the trial in the first year were approached to participate in a semi-structured interview that included questions about their school and the use of APP. One head teacher did not take part in the interviews and did not respond to email or telephone follow up contact and another declined and explained that they had nothing to say due to limited involvement with the study.

These semi-structured interviews were conducted by experienced qualitative researchers by telephone using three standard questions embedded within a larger interview about their experience of the research trial. Questions addressed how teachers determine APP levels, what systems of training and moderation are in place, and what the resulting APP scores were used for in their school.

Data from the head teacher interviews were analysed using a process of deductive thematic analysis as outlined by Braun & Clarke (2006). Two researchers (BN, RM) coded the data in line with the a priori framework and the codes were used to clarify concepts, map the range of the phenomena, create typologies and explore associations.

Results

1) WIAT- APP comparison

Insert Figure 1 here

The mean (SD) age of the participating children was 8.0 (1.1) years and thirty nine were boys *(Table 3)*. Fourteen (19%) of the children had special education needs. One of the 11 teachers was male; the median (IQR) age of teachers was 33 IQR (26 to 42). The median time that they had been teaching was 6 years (1 to 10) and 5 of the 11 teachers had taught for fewer than 5 years.

Insert Table 3 here

All 72 children had data provided on the APP. One child was missing on WIAT literacy and two children were missing on WIAT numeracy. Table 4 shows the ability the children assessed according to the APP scores supplied by the teachers.

Insert Table 4 here

Data analysis revealed the Pearson correlation between the APP and the WIAT was 0.73 (95% CI: 0.59 to 0.82; $p < 0.001$) for literacy and 0.12 (95% CI: -0.12 to 0.34; $p = 0.33$) for numeracy.

Insert Figure 2 here

Further analysis (Table 5.) revealed that there were no notable differences between the correlation between APP scores and the two mathematics subtests administered. There were higher correlations between the reading comprehension subtest and APP literacy scores than the word reading or spelling subtests. The numbers of children completing the Reading comprehension subtest was reduced because the age of administration for this subtest is higher.

Insert Figure 5 here

Tests of interaction generally revealed little evidence that either of these correlations differ between children with and without special education needs, nor between children taught by teachers who had practiced for less than 5 years and those who had taught for more than 5 years. There was weak evidence that that the correlation (p=0.08) for literacy differs between children with special education needs (0.50; 95% CI: -0.06 to 0.83) and those without (0.72; 95% CI: 0.57 to 0.82), a larger study would be required to investigate this conclusively.

2) Head Teacher Interviews

Thirteen head teachers commented on how APP is used in their schools. Data is outlined in relation to the initial a priori thematic framework, as no data were identified that did not fit within it. The framework (as informed by the questions asked) is outlined below:

How do teachers come to conclusions regarding APP levels?

Most head teachers reported that teachers developed APP scores through a process of ongoing 'day to day' assessments throughout the year, which for most included observations of the child's performance in school and professional judgement.

'Everything, so it would be observations, it would be their annotated planning… it might be some summative assessments, it would be their marking, it would be the information they've gathered

through questioning and the range of assessment for learning strategies that they've employed, to build a picture of where the children are.' HT 06

Two schools noted that they used optional standardised tests (NC) to inform their APP levels and other schools noted that they would use formal tests in order to inform their judgements of APP levels. One school said that their process was very formal, using a selected range of written work and predefined assessments to develop a 'portfolio'.

'The teachers take a number of pieces of writing over a term, that's what they're basing APP assessments on, rather than it being one particular piece, they take three or four out of the term, so we're building a picture. They also take optional SATS through key stage 2.' HT 05

Two schools said that they used a computer programme to do this but both noted that this was never used in isolation and some also commented on the 'idiosyncrasies'/ individual differences between the children being assessed and criteria. Two heads also noted teachers using the APP grids/ descriptors to help them form judgements and a further two schools talked about the APP criteria solely in relation to a children who were deemed to be struggling.

Teachers were reported to observe any change in the child's performance as requiring a potential change of APP score, however there were difficulties related to how consistent this change needed to be in order for the child to progress upwards.

What training/ moderation occurs in relation to APP?

Seven of the head teachers reported specific moderation meetings at least once every half term to ensure consistency of APP scores. Of these, three schools also moderated with another school, one of which was a secondary school. Other methods of moderation were across teams within schools, or across year groups or as a whole staff team. One school said that they rarely used moderation. The other five schools mentioned more informal moderation that occurred in meetings convened for other reasons, such as continuing professional development. One schools also appointed teachers

with a mentor who they could consult regarding APP. There was quite often a system within schools of senior teachers, such as leaders across the year, the curriculum or assistant heads as having a core role in 'checking'/ moderating APP levels.:

'you unpick what the statements mean and then you think about what it looks like in your child's work and have you thought the same. And then you get one- to – one support where they would moderate their judgements with a member of the senior leadership team who checks them'

HT06

This process of taking specific examples and extracts of work to moderation meetings to check others' judgements against your own was very common in the processes outlined. Benchmarking was also noted as another important feature of moderation. Moderation happened on an individual/ one-to-one basis or on a group level, this differed between schools.

No school said that all teachers had been trained in using APP. Two schools said that they received training from 'curriculum leaders within the school' and a further two heads said that two or three members of staff had been on training when the approach was first introduced and then brought this training back to the school.

'Well in the normal raft of training that we do, I mean curriculum leaders provide in-house training on the processes. We've also had three members of staff when the APP came about, we had three members of the assessment team, myself and the two key stage leaders at that time attended those courses and we fed back to staff and then it's an ongoing updating process' HT11

How are APP scores used in schools?

The APP scores were used differently amongst schools, only ten heads responded to this question, and of these eight heads said that they use APP scores formally to monitor children's progress within

school. For six of these heads this progress monitoring would help trigger thoughts about additional support that the child may need within school:

'First of all, identifying our key children who with extra support would be able to move on.'

HT12

Other uses of the APP included;

'Reporting back to parents, analysing data, checking on children's progress, reporting to other teachers so you know where they are, OFSTED.' HT10

Some schools were very clear that the APP scores derived were only part of an overall 'teacher assessment'.

Head teachers also reported that APP have both a summative and a formative purpose. For example, APP was used to inform judgements of a child's progress throughout the school year. For some schools this was very clearly being tied to teacher performance and accountability. Progress was a core feature of the use of APP where scores were then abstracted to national performance standards and from this a point-score change in progress noted for each child. This point-score enabled the teacher to map the child's progress.

'We will look at each child, the starting point, progress they've made and if they've made progress or above, but any child that hasn't made progress that's expected, I'll challenge that and we'll talk about what we're doing to ensure that they, they're making the expected progress.'

HT3

Discussion

We compared longitudinal APP assessments completed in February 2013 with performance on standardised tests undertaken at the same time. The correlation between assessments was high for

literacy but much lower for numeracy, and did not vary according to subtest despite thoughts that the mathematical reasoning subtest may map less well onto APP criteria and therefore that correlations for numeracy might be greater for numerical operations. On the literacy subtests, the highest correlations were with the reading comprehension subtest, which mapped on clearly to the APP criteria, while the low correlations between literacy and spelling on the WIAT may be expected given the lack of emphasis on spelling in the APP criteria.

There still remains the question regarding why there was such a low level of agreement between the APP and WIAT scores for numeracy. One explanation may be a greater tradition of negotiation of achievement standards in literacy. Another might be that the most recent WIAT (WIAT -11) was normed in 2004 in the UK, and teaching practices may have changed since.

The timing of assessments might have contributed to the findings. The WIAT assessments were conducted mid-way through the academic year, to ensure that the teacher would have had adequate time to observe and get to know the child, but to exert as little disruption on the school's programme and the burden for teachers as possible and schools. Head teachers noted that most assessments of progress were made at the end of the year to compare to pre-defined targets. It may be, therefore, that the teachers provided APP scores based on more tentative judgements than they would have at the end of the academic year. Only two schools said that they used formal assessments (optional NC tests) to inform the APP level, so it seems unlikely that lack of access to formal assessment results impacted greatly on the data presented, and even if it had, it would not explain the different levels of agreement obtained in respect of literacy compared to numeracy.

The teachers union, NASUWT (2010), issued a briefing note that recognised the potential benefits of the APP framework and also outlined examples of 'poor' APP practice, as reported by some of their members. Many of the examples of poor practice were explicitly contrary to the guidance for using APP, for example, using APP in addition to existing assessment arrangements, using APP more frequently than termly or 6 monthly, too much in-school moderation, undertaking specific

assessment activities to generate portfolios and its application without support and adequate preparation. It was unclear how widespread these deviations from recommended application of the APP were. From the interviews with head teachers, it became clear that schools use and moderate APP very differently and therefore variation from recommended practice may be widespread. The fndings also shows how APP can be diverted from its intended methods and purpose.

It is also important to recall that the original purpose was that APP be used for formative purposes but was being used summatively. As our heads noted in this study, results from the APP are used externally, for parents and Ofsted, and internally to inform decisions about access to further support for children and to monitor teacher performance. Ofsted (2011) saw APP as having both formative and summative roles, teachers and schools may only focus on their summative nature. Indeed, Ofsted's (2011) survey reports some disappointment about the lack of integration of assessments into steering the learning of children (i.e. in formative assessment). Part of the historical difficulty with teacher-led assessments is the lack of clarity about their role and purpose. Assessment results have been shown to have a significant impact on a young person's academic life (Broadfoot & Black, 2004) and also on their sense of self-efficacy, motivation, enjoyment of school and willingness to learn other than for assessments (Harlen, 2005). The impact of relative age, that children who are older for their year achieve more at school and have a greater sense of self-efficacy and a more internal locus of control (Crawford, Dearden & Greaves, 2013), is hypothesised to be partly due to the impact of assessments which place them as more able than younger people in their year. Therefore, assessment results have the potential to significantly impact a child and their development. This is important to note as the qualitative interviews highlight that there is a discrepancy between the intended aim of the assessment according to APP and what the assessments are actually quantifying. It may be that teachers' assessments do not or should not aspire to provide a judgment on a child's ability in relation to others of the same age. It may be that the goals of these assessments vary or should vary, for example in reporting the child's ability in relation to specific learning goals.

Alternatively, the emphasis may need to shift to helping make teacher assessments more accurate. There might be individual factors to take into account. Some research addresses how teachers develop judgements about children (Allal, 2013; Wyatt- Smith, Klenowski,& Gun, 2010; Wyatt- Smith Klenowski, 2013) or how biases related to gender, incorporation of non-relevant behaviour by teachers, special educational needs and a discrepancy between the child's verbal and other abilities can impact (Harlen, 2004 a & b). Wyatt- Smith & Klenowski (2013) showed that teachers had their own criteria that they drew on for assessment and that comprehensive specification of criteria and standards were only part of what informed teachers' assessments.

Literature suggests that in order to enable teacher assessments to more closely represent what is wanted by national bodies, teachers must have clear guidelines (Harlen, 2004 a & b). Harlen's (2005) systematic review reported that teacher assessments are made more reliable and valid if they have clearly defined, dependable criteria. It may be that teachers in this study had more difficulty in applying numeracy APP criteria. The science APP criteria had already been rejected as they were not easily applied in primary schools, which suggests that teachers involvement in the development of any criteria might assist their implementation. Given normal variations in child development, this is a substantial task.

It may be that a different method of comparison can be used. It may be unfair to assume that teacher's assessments are comparable to norm-referenced assessments. It may be unrealistic to assume that teachers should be able to accurately assess intellectual functioning, especially given the significant demands placed upon them. In this case a review of the scope, function and purpose of teacher assessment would need to be completed.

Ofsted (2011) concluded that the APP initiative tended to help schools strengthen their assessment practices to a greater extent when it was part of a strongly led, whole school vision about teaching, learning and assessment. The vast majority of head teachers reflected this opinion and had indeed developed a wide-level approach to teacher assessment through APP, either at a year group or

whole school level. Integral to this was a system of moderation and training within schools, which is an area worthy of future research development. How schools conduct moderation and what they do would seem to be very important in ongoing debates about teacher assessment, especially given the impact of the results of assessments on schools, teachers, families and children. These will remain significant issues even though APP teacher assessments are being phased out in the UK. Policy changes might temporarily change teacher practices, but the status of teacher assessment in the estimation of learner attainment and progress will continue to be an important matter.

Acknowledgements

This report presents independent research funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health in England.

National Institute of Health Research, Public Health Board Grant number 10-3006-07

References

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice,* 20:1, 20-34

Black P., Harrison, C, Hodgen, J, Marshall, B., Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning?, Assessment in Education: Principles, Policy & Practice. 18 4, 451-469

Braun, V., Clarke, V., (2006). "Using thematic analysis in psychology". *Qualitative Research in Psychology* **3** (2): 83

Broadfoot, P., & Black, P. (2004): redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice*, 11:1, 7-26

Clay, M, M. (2006). *Observation Survey of early literacy achievement.* **Heinemann Educational Books**

Crawford, C, Dearden, L & Greaves, E. (2013). *When you are born matters: evidence for England*. Institute for Fiscal Studies

Department for Child, Schools and Families (2009) *Evaluation of Making Good Progress pilots*; RB184. London: DCSF.

Department of Education (1987) Report on Task Group on Assessment and Testing. London: DES.

Department for Education. (2010). Progression 2010 to 2011: advice on improving data to raise the attainment and maximise the progress of learners with special needs. Department for Education.

Department for Education (2013). Freedom of information request 2013/0017677. 11th April 2013.

Durant, D. (2003). A comparative analysis of Key Stage tests and teacher assessments. *Paper presented to the British Educational Research Association Annual Conference. Heriot- Watt University Edinburgh*

Elliot, C, D. & Smith, P. (2011). *British Ability Scales - Third Edition (BAS-3).* GL Assessment

Ford, T., Edwards, V., Sharkey, S., Ukoumunne, O., Byford, S., Norwich, B., Logan, S. (2012).

Supporting teachers and children in school: the effectiveness and cost effectiveness of the incredible years teacher classroom management programme in primary school children: a cluster randomised controlled trial, with parallel economic and process evaluations. *BMC Public Health*. 12. 719.

Harlen, W. (2004a) A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes (EPPI-Centre Review), Research Evidence in Education Library, issue 3 (London, EPPI-Centre, Social Science Research Unit, Institute of Education). Available http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=116
Downloaded 8th May 2013

Harlen, W. (2004b) A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes (EPPICentre Review), Research Evidence in Education Library, issue 4 (London, EPPI-Centre, Social Science Research Unit, Institute of Education).
http://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=OdG3t7IX5EY%3D&tabid=119&mid=925.
Downloaded 8th May 2013

Harlen, W. (2005). Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal.* 16 (2). 207-223

Johnson, S. (2013) On the reliability of high-stakes teacher assessment. *Research Papers in Education*. 28, 1, 91-105

Moseley, D. (2008). *Word Recognition and Phonic Skills 3rd Edition Manual.* Hodder Education.

NASUWT (2010) Assessing Pupil Progress. Retrieved at
http://www.nasuwt.org.uk/InformationandAdvice/Professionalissues/APP/#

NUT (2009) Assessing Pupils' Progress: manageability. Retrieved from
www.teachers.org.uk/files/APP-manageability-17Mar09update.pdf

Ofsted, 2011. The impact of 'Assessing pupils' progress' initiative. Downloaded 11th March 2013: Available at www.ofsted.gov.uk/publications/100226

Qualifications and Curriculum authority. (2008). Assessing pupils progress: Guidance for planning and supporting in-school standardization and moderation. *National Strategies*

Qualifications and Curriculum authority. (2009). Get to grips with assessing pupils progress. Department for Children, schools and families.

Qualifications Curriculum authority. (2010a). Assessing Pupils Progress: Assessment Criteria:

Number and algebra/ Using and applying mathematics Shape, space and measure Handling data. Qualifications Curriculum authority. Department for Children, schools and families. http://nationalstrategies.standards.dcsf.gov.uk/node/20683 28/03/2011

Qualifications Curriculum authority. (2010b). Assessing Pupils progress: assessment focuses and criteria: APP Speaking and listening/ writing/ reading. http://nationalstrategies.standards.dcsf.gov.uk/node/20683 28/03/2011

Wechsler, D. (2005). Wechsler Individual Achievement Test 2nd Edition (WIAT –III). London: The Psychological Corp.

Wilkinson, G., S. (1993). *Wide Range Achievement Test, Version 3.* Psychological Corporation,

Wyatt- Smith, C, Klenowski, V, & Gun, S., 2010. The centrality of teachers' judgement of practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice.* 17: 1, 59-75

Wyatt- Smith, C & Klenowski, V. (2013). Explicit, Latent and meta criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice,* 20 1, 35-52

Table 1. Other tests considered

Measure	**Author**	**Age Range**	**Areas assessed**
WRAPS (Word Recognition and Phonic Skills 3rd Edition)	Moseley (2008)	4.6-0	Word recognition and Phonic skills
British Ability Scales (BAS-3)	Elliott & Smith (2011)	2.6-11	Word reading test; word recognition and verbal reasoning and knowledge
Observation Survey of early literacy achievement	Clay (2006)	6 years	Letter Identification, Word Test, Concepts, About Print, Writing, Vocabulary, Hearing and Recording Sounds in Words Text Reading
Wide Range Achievement Test (WRAT – 3)	Wilkinson (1993)	5-75	Word recognition and writing
Wechsler Individual Achievement Test. (WIAT – II)	Wechsler (2005)	4-16	Reading, Numerical Attainment, Language attainment, oral expression. *Note:* Superseded the WORD and the WOND

Table 2. Summarised APP criteria and areas assessed by WIAT

App Criteria	WIAT subtests selected
Numeracy a) Number and algebra (including calculation) b) Using and applying mathematics. Describing shapes, measurement, angles, space. Handling data; using tables graphs.	*Numerical operations*. Evaluates the ability to identify and write numbers, counting, calculation and solving equations. *Mathematical reasoning*. Ability to identify geometric shapes, solve single and multi-step word problems, interpret graphs and identify mathematical patterns.
Literacy a) Speaking and listening; talking to others exploring ideas, listening, making comments asking questions, using imaginative play and understanding meaning. b)Reading; decoding text to read, understand and select relevant information, interpret/ infer information or ideas from stories. Identify and comment on the structure of text including grammar and punctuation. Relay the writer's viewpoint and relate to context/culture/ history. c) Writing; vary sentences for purpose, clarity and effect, write with technical accuracy. Use correct spelling.	Not covered *Word Reading.* Identifying letters of the alphabet, sounds in words and accuracy and speed of reading aloud. *Reading comprehension*. Evaluates reading written instructions in the classroom. Matching a written word to its picture. Reading passages of text and answering content and comprehension questions. *Spelling*. Evaluates the ability to spell and write dictated letters, letter blends and words.

Table 3. Mean scores and standard deviations

	Mean Score	Standard Deviation
APP Literacy	13.0	5.6
APP Numeracy	13.4	5.6
WIAT literacy	96.9	16.3
WIAT Numeracy	100.0	11.6

Table 4. The number of children performing below/at/above average by age at 1st September in the sample

	N (%) Below Average	N (%) Average	N (%) Above Average
Age 5 (N = 19)	14 (74%)	4 (21%)	1 (5%)
Age 6 (N = 7)	6 (86%)	0 (0%)	1 (14%)
Age 7 (N = 31)	2 (6%)	12 (39%)	17 (55%)
Age 8 (N = 15)	2 (13%)	7 (47%)	6 (40%)
All (N = 72)	24 (33%)	23 (32%)	25 (35%)

Table 5. Correlations: WIAT and APP including subtest analysis

Outcome	Aggregate Measures	Correlation	Lower bound	Upper bound	N
APP literacy	Versus WIAT literacy	0.73	0.59	0.82	71
	Versus WIAT word reading	0.06	-0.17	0.29	72
	Versus WIAT reading comprehension	0.62	0.43	0.76	59
	Versus WIAT spelling	0.05	-0.18	0.28	71
APP numeracy	Versus WIAT numeracy	0.12	-0.12	0.34	70
	Versus WIAT numerical operations	0.18	-0.06	0.39	71
	Versus WIAT mathematical reasoning	0.04	-0.20	0.27	71

Figure 1

Figure 1 Flow diagram to illustrate the qualitative and quantitative methodology

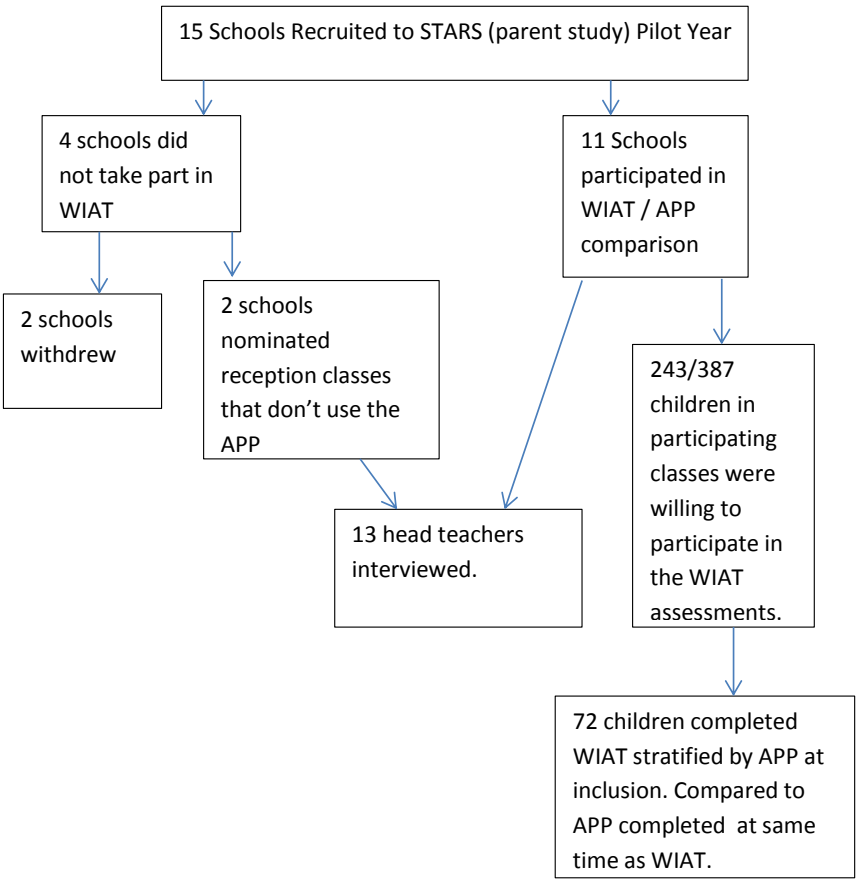

Figure 2. Scatterplot of WIAT Scores versus APP score