Appendix 3 Adequacy of the Key Stage Test Assessments

There is some concern about the adequacy of the Key Stage test assessments. The use of the Key Stage tests as a reference for teacher assessment assumes that the tests are of adequate validity and reliability. Stobard (2001) considers the validity of the Key Stage tests and concludes that

> Even though the test development process is exemplary (Wiliam, 1999), the validity of their use as the key measure of year-on-year changes in national standards is questionable. …. (some) fluctuations may best be explained by problems with test equating rather than dramatic changes in standards in particular subjects. The consistency of Teacher assessment scores supports this interpretation.(Stobard, 2001, p. 32)

Stobard cites the official data on tests and teacher assessments for Key Stage Three over the period 1997 to 2001 (Stobard, 2001, p.33, Table 1) which show a pattern of approximate comparability. Thesis Figure 4.1 shows an equivalent trend for Key Stage 2. The percentages of students reported at level 4 in Figure 4.1 or higher from the two assessment sources differ from each other by as much as 6 percentage points (science in 2000) but most often differ by 3 points or less. There is greater variability in the test results around the general trend than the teacher assessments, although both follow approximately similar trends. Stobard argues that the greater stability of the teacher assessments is an indicator of the lesser validity of the test results (p.32).

Uncertainty about the consistency of the KS tests over time is shared by others. Tymms (2004) raises a range of concerns about the adequacy of the Key Stage tests as measures of performance of students over time. He queries whether cross-sectional cohorts of students are moving up the levels scales in English and mathematics over a sequence of years at the rate indicated by the official figures. He identifies two periods of change in student scores: 1995 to 2000, where scores rose rapidly and from 2001 to 2004 where scores appeared to hardly change. Student time series scores are represented at Key Stage 2 by the rather gross measure of 'percentage at or above level 4' in the Key Stage 2 tests. This indicator moves from being about average (of the order of 50% of students above and below the position in 1995) to the indicator showing over 70% of students above the position by 2003. Based on eleven data sources Tymms is concerned that the statutory tests might overstate the improvement from 1995 to 2000, and may have underestimated the rise from 2001 to 2004.

Part of the reason for the concern is the process for setting cut scores each year. Tymms citing Quinlan & Scharaschkin (1999) summarises the four processes applied as; "marker opinion, professional scrutiny of the test papers (Angoff techniques), earlier use of the live test and the employment of an anchor test" (Tymms, 2004, p. 489).

Two difficulties are identified by Tymms. The first is that the cut scores were required to correspond to a 'mark', that is an integer. He explains that "a change of one mark in 1996 would have made about a 1.4% difference in the proportion of students being awarded level 4 or above" (Tymms, 2004, p. 484). The effect in 2000 is estimated to be 1.8%.

The second difficulty is that prior to 2001 standards were only equated from one year to the previous year. Tymms argues that this allowed drift in relative standards due to the compounding of error over time. Post 2000 the equating process applied over multiple years, and "may well be the reason for the abrupt changes between Phases 1 and 2" (p. 490). (Phase 1 is the period 1995-2000; Phase 2 is the period 200-2004). Tymms concludes that the use of the tests in the period 1995 to 2004 for "monitoring standards over time…has failed for a number of reasons" (p. 492). As a result of this analysis some caution is needed in accepting the quality of the test data, especially when it is compared with, or used as a standard for,

teacher assessment. Unlike Stobard, Tymms has not used the teacher assessments as a possible guide to the 'real' change over the period.

Reference to the regular test statistics (Test Statistics, 2007) published on the QCA/NAA website provide an additional insight into the test development process. Tests in a given subject are developed by individual contractors (NFER, Edexcel as examples) without necessarily continuity of contract from one year to the next. The publicly reported analyses are very simple (Cronbach's alpha, percentages correct for each item in the trial samples, score ranges for mapping to levels i.e. 'level threshold tables'). The statistics summaries indicate that whatever analyses are made in test development, the allocation to a level is based on test marks as reported by Tymms. No item maps or indication of the spread of item difficulties for those items around the level boundaries are given. Individual student reports to parents are level only, not scores (*End of key stage 2 pupil results proforma*, 2008). The marked scripts are returned to schools, so schools are informed of 'marks', but the computer file/report provided with them indicates only a level in each subject (*Assessment and reporting arrangements* (ARA) 2009). Head teachers must report the result to parents 'within 15 school days of the head teacher receiving it' (ARA, 2009, p. 80). From the evidence of Tymms, Quinlan & Scharaschkin, Stobard and the QCA publications, the Key Stage tests and teachers assessments are reported to schools on a very broad scale. The subsequent summaries at Local Authority and national level used in times series reports are even broader, concerned mainly with the percentage of students that are at or above a particular level, depending on the Key Stage. Given the process, a general impact of measurement error is expected around the threshold for the appropriate level in each Key Stage.

Based on the evidence above, using the test data as the 'assumed' best possible independent estimate of a student's developmental position on the levels scale is problematic. Mismatch of teacher and test data would not necessarily indicate any inadequacy in the teacher assessment but may reflect some general looseness in the test data, even given the broadness of the level scale. However being aware of the patterns of the two assessment processes over time and the persistence of these patterns by subject, provides some hints as to the relationship.