

Report from the Working Group on the InCAS Errors of October 2009

Prepared by **Professor John Gardner (Queen's University)**

for the **Department of Education**

15-April-2010

Contact:

Professor John Gardner

School of Education
The Queen's University of Belfast
69 University Street
Belfast BT7 1HL

Tel: 028 90 975017
e-mail: j.gardner@qub.ac.uk

Working Group on the Impact of InCAS Errors¹ in October 2009

Summary

1. There is little doubt that the errors communicated to schools during October 2009 have had considerable impact on the confidence of school principals, teachers, parents and pupils in the results of InCAS assessments. The first point to be made in this report is that the error situations did not arise as a result of any problem specific to the design of the InCAS system or its functions. They had been introduced into the computer coding in 2009 and did not exist during the two previous years of successful InCAS usage. The errors were entirely human-based with simple fixes that involved small corrections in the computer software that carries out the analysis of pupils' scores. As such they are unacceptable in terms of quality of service and should not have happened.
2. The first error occurred in the General Maths scores of the 94,439 pupils who took this assessment in 886 schools. The results of 34,271 pupils were potentially affected before the error was fixed. These were pupils who attend the 328 schools that had already uploaded their results and downloaded the analyses before the error was corrected. Of these, 79 schools had begun to communicate the erroneous age-equivalent scores to the pupils and parents.
3. Once the error was detected CEM corrected the scores and analysed them to establish the extent of the errors across the Year groups. This analysis showed that 60% of the affected pupils (20,472) had errors in their age-equivalent scores of 1-6 months (with another 3,129 having errors of less than one month). Of the remainder, 7% (2,450) had errors of more than one year. The large majority of the pupils therefore had relatively small errors in terms of age-equivalent scores.
4. Even without a human error situation, it is not unreasonable to consider that parents will ordinarily be concerned by an age-equivalent score below their child's real age, thinking, as they might, that their child is falling behind. Their child might indeed be falling behind but teachers are trained to explain that the InCAS scores are estimates that can be influenced by various circumstantial factors. The teachers are required to use the scores to augment judgements of strengths and weaknesses that are based on their own, more extensive knowledge of the children.
5. In the circumstances of human error, with incorrect age-equivalent scores communicated to parents (and their children) and with teachers and principals required to explain the situation and its implications, it is reasonable to assume that considerable confusion and frustration would have developed. The most substantial impact would have been felt by parents (and their children) and the teachers who had to cope with large discrepancies in the age-equivalent scores that had already been communicated. But many principals and teachers, in the 79 schools mentioned above and in the 83 others that had begun to prepare for meetings with parents, would have had the added burdens and frustrations of repeating the analyses, interpretations and print-outs.
6. These various figures indicate that overall it was a small proportion of schools, parents and pupils who experienced direct impact of substantial errors in communicated scores. However, the media coverage of the errors would have raised concerns and shaken confidence in all schools and would have extended to many more parents with primary school children than those directly affected.

¹ Details of the InCAS system and various aspects of the error events are available at http://www.nicurriculum.org.uk/key_stages_1_and_2/assessment/InCAS/

7. The situation prompted immediate action at the highest levels, supported by public statements from the Minister and the Council for Curriculum, Examinations and Assessment (CCEA). Assurances were demanded from and given by the supplier, the Centre for Evaluation and Monitoring, University of Durham (CEM) that the error was corrected and that the whole system had been thoroughly checked. However, the second error came after these assurances and immediately inflamed the concerns and frustrations of the schools. Even though this error, in the calculation of standardized scores, did not affect the statutory reporting to parents, it did affect all aspects of the various² InCAS assessments that are designed to inform teachers' judgments about their pupils' strengths and weaknesses. Again it was relatively easy to correct but the impact of the error, coming on the back of public assurances of all results now being robust and dependable, was something of a public relations disaster and a major blow to schools' confidence in the InCAS system.
8. These comments (and the detail of the report below) present a gloomy but fair assessment of the impacts of the impact on schools', parents' and pupils' confidence in InCAS. However, in the manner of a 'silver lining to every cloud', there are arguably two significant benefits arising from the situation.
9. The first is that teachers, parents and pupils now have good reason to accept that assessment results can be subject to error, albeit usually as a natural degree of in-built uncertainty rather than the more explicit conditions of human error experienced here. For many more people, interpreting scores may well in the future be a more measured process of being aware of potential errors (either the natural levels of uncertainty or human mistakes) and being vigilant about checking them.
10. The second benefit, albeit a little belated, is the recognition beyond any doubt that despite the InCAS assessments not being designed to be 'high stakes', there is an exceptionally high perception of their importance among parents and schools (teachers and principals). It follows that greater attention is now being paid to quality control in all aspects of the assessment process and reporting. DE and CCEA have initiated key areas for action including ensuring that the quality control processes are sufficient to accept assurances from CEM with confidence, and identifying recommendations for rebuilding the confidence that existed in InCAS before the recent errors occurred.
11. In this latter respect this report sets out a number of recommendations aimed both directly at the error situations that arose and on wider aspects of computer-based diagnostic assessment in primary schools.

Recommendations

12. It is recommended that:

1. steps are taken to emphasize to parents that the primary purpose of the InCAS assessments is in contributing to diagnostic support for pupils' learning – and not for comparing pupils.
2. the trialling of items should be separate from formal use of the assessments, perhaps being developed in collaboration with schools that volunteer to participate.
3. if the formal use of the assessments must continue to incorporate trial items, this should be notified to pupils and teachers so that they know the additional items do not contribute to the assessment.

² Only General Maths and Reading assessments (English and Irish Medium versions of both) are required to be reported to parents. Four others: Spelling, Mental Arithmetic, Developed Ability and Attitudes are for diagnostic purposes only.

4. arrangements are made to test the performance of the InCAS data processing before general release of analysed age-equivalent and standardized scores. It would not be practical or sensible to do this for all schools as the response times for analysed data would probably be seriously extended. However, it could be accomplished by collaborating with a selection of volunteering schools which provide their data early in the autumn term and contribute to the necessary checking and feedback.
5. the annual summer term briefings of relevant school staff on the upcoming autumn assessment period should include information on the measures taken to ensure the system is robust and dependable.
6. InCAS-related training should increase the emphasis on the interpretation and sharing of InCAS results with parents.
7. CCEA and DE take steps to ensure that the CEM quality control procedures are regularly and independently reviewed to reduce the potential for recurrences of errors of the types experienced in October 2009.
8. meetings with principals are convened in the third term of 2009/10 to set out and give assurances on the steps being taken to secure the reliability of the system in the future.
9. all schools should be asked to give specific feedback (perhaps through the annual evaluation surveys) on the diagnostic use and performance of the assessments.
10. research should be commissioned into the extent of usage of commercial practice tests by schools and parents.
11. research should be commissioned into how teachers and principals use InCAS scores for diagnostic, planning and evaluative purposes.
12. CCEA and DE seek to recoup the costs of substitution cover from CEM.
13. the legislative dimension of computer-based assessments should be reviewed.
14. the sizes of item banks are considerably increased in order to improve diagnostic potential and to maintain credibility by reducing both the risk of memorization and the effectiveness of practising as a means to improving performance levels.
15. differential item functioning analyses are carried out for sub-groupings within the pupil population including gender, second language, special needs, socio-economic status and minority ethnic groupings.
16. InCAS consider offering standardized scores on the basis of raw scores and that these standardized scores should be referenced to both Year cohorts and age.
17. CCEA and CEM agree a policy on standardization as soon as possible.
18. CEM institute appropriate modifications to ensure that information on the various types of inappropriate access to and usage of InCAS is available and traceable.
19. a review of ICT resources for InCAS administration is carried out to identify schools needing support.

Introduction

13. During October 2009, two errors were identified in the way scores were calculated for the compulsory Years 4-7 InCAS assessments in primary schools. These were accompanied by a considerable period of confusion and frustration for schools and parents, fuelled by media interest and formal questions at the highest levels of government. Once the errors were rectified, the Department of Education (DE) in conjunction with the Council for Curriculum, Examinations and Assessment (CCEA) focused on five key areas for action, one of which was 'confidence re-building'. To this end a working group was appointed (for membership, please see Appendix 1) with a remit to:

"... provide an assessment of the impact of the two errors on school confidence and make recommendations to DE and CCEA (and ESA when established) on actions that should be taken before autumn 2010 to improve confidence among schools, teachers and parents in the 2010 InCAS assessments".

14. The Working Group was also empowered to:

"... provide views or recommendations on wider aspects of computer-based diagnostic assessment in primary schools here".

15. In carrying out its work:

"... the working group must ensure that [its assessment, recommendations and views] reflect the current legislative requirement for computer-based diagnostic assessment to be carried out in the autumn term and the outcomes reported to parents".

16. The Group considered the issues over the short time-frame 11th February to 31st March, 2010. The Group based its report on the impact of the errors on written and oral evidence presented to them. This comprised various papers from DE, CCEA and CEM; written comments from teachers and principals; anonymized correspondence from schools, teacher unions, parents and the media, and oral inputs from a variety of principals, teachers, parents and governors with whom the members of the group consulted. The view of the errors and their impact reported here are therefore considered to be reasonable interpretation of the situation that arose and its effect in undermining confidence in the system.

17. The report that follows begins with a detailed assessment of the impact of the errors and recommendations on how to improve confidence. This is followed by wider views and recommendations relating to the system of computer-based diagnostic assessment in Northern Ireland primary schools.

Assessment of the Impact of the InCAS Errors in October 2009

18. The evidence available to the Working Group strongly suggests that negative impact from both error events was felt by several groups: pupils, parents, teachers, school principals and both the Council for Curriculum, Examinations and Assessment (CCEA) and Department of Education (DE). The matters were treated seriously and urgently by the managers of the system, CCEA, and by the suppliers, the University of Durham Centre for Evaluation and Monitoring (CEM), and considerable staffing resources were allocated to rectify the problems within days of identifying the causes.
19. For the most part it was the General Maths error that had the potential to upset the majority of stakeholders: pupils whose progress was wrongly assessed, parents who were misinformed, teachers who felt exposed and let down by a system they are mandated to use, and principals who had to manage the process with staff and parents.
20. In contrast, the standardization error impacted almost exclusively on schools and teachers, who should be able to use the standardized results to inform their knowledge of how their pupils and classes are performing. This is not to say that it was a lesser error. Both errors are unacceptable but in effect the standardization error and its media attention added considerably to a level of confidence that was already badly shaken by the impact of the first error. A second error in quick succession, especially following unqualified apologies and absolute assurances of reliability, was almost particularly damaging.
21. In the sections that follow the primary focus is on the General Maths error as its reach covered a wider range of people. The standardization error is dealt with separately as its impact was more specific.

Pupils

22. It is not possible to judge whether there is any lasting impact on the pupils whose incorrect General Maths age-equivalent scores were communicated to them and their parents. Of the 94,439 pupils assessed in 886 schools during the first term of the 2009/10 school year, the results of 34,271 pupils were potentially affected before the error was fixed. These were pupils who attend the 328 schools that had uploaded their results and downloaded the analyses before the error was corrected. Of these, 79 schools had already begun to communicate the erroneous age-equivalent scores to the pupils and parents.
23. Once the error was detected CEM corrected the scores and analysed them to estimate the extent of the errors across the Year groups. This analysis showed that 60% (20,472) had errors in their age-equivalent scores of 1-6 months (with another 3,219 having errors of less than one month). Of the remainder, 7% (2,450) had errors of more than one year.
24. The error had arisen from the inclusion of a set of experimental questions placed at the end of the assessment. The score processing software should have excluded these items but the specific code was missing. The offending items were targeted at the higher levels of difficulty and meant in effect that able or older pupils would have been most likely to reach these items. As a result, the impact of the error was largely restricted to the scores of Year 7 pupils. Of these pupils, 45% of the age-equivalent scores were identified as having errors of five months or more and 25% had errors of 11 months or more³. The majority of these errors registered the pupils' age-equivalent scores as higher than they should have been.

³ Figures calculated from CEM paper: *InCAS General Maths Scores, 12th October 2009*

25. The analysis of corrected pupil data from the other Year groups show that the numbers experiencing errors of five months or more were 23% for Year 4, 28% for Year 5 and 30% for Year 6. Although it seems improbable, especially for the Year 4 pupils for example, that these large proportions of pupils progressed to the higher difficulty trial questions, this is the explanation offered in the reports for these relatively high discrepancies. For errors in excess of nine months the proportions of pupils in each Year group were, as might be expected, very small: 2% for Year 4, 3% for Year 5 and 4% for Year 6.
26. The effect of the error on age-equivalent scores was primarily to inflate them, giving results with higher age-equivalents than should have been the case. For example, a student with a true age of 11 years 5 months may have been attributed an age-equivalent score of 12 years 2 months.
27. Some evidence of pupil upset, for example concerns about whether they would have to repeat the assessments, is available but it could be argued that an error which inflates scores is less directly upsetting to a pupil than one that suggests they are performing at less than the average for their actual age. However, any upset would depend on whether the subsequent correction took them below their own expectations and the score suggested for their age-equivalent peer group. More importantly, however, any error has to be seen in terms of the wider context of the pupils' performance being evaluated.

Survey of Pupils' Views

28. CCEA conduct a survey of pupils', teachers', principals' and parents' views on InCAS each year and in the first term of 2009/10 11,117 Years 4-7 pupils completed on-line questionnaires⁴. As with all of the surveys there was no mention of the General Maths error (nor was the standardization error mentioned in the teachers' and principals' instruments) but 81% (9,024) of the pupils answered 'No' to a question asking whether they found any difficulties in using InCAS. Again it is not clear from the interim report whether the remaining 19% (almost one in five) responded 'Yes', that they did find it difficult, or if a proportion of these had not answered the question. Some 31% expressed the view that the assessments could be improved but at the time of writing the analysis of any narrative expansions of this view was not yet available. The instrument did not ask whether the pupils felt the InCAS results matched their perceptions of their own abilities in General Maths and Reading.
29. InCAS is explicitly designed to be diagnostic and not high-stakes; i.e. not having any consequence in terms of life-choices as a result of the score. However, practice is somewhat different from the theory and reports from schools and parents do refer to the importance attached to doing well in the assessments. The angry brouhaha over the error events is testimony to the degree of importance attached to any form of pupil assessment or testing that goes wrong.

Parental Expectations and Its Influence on Pupils

30. It is natural for parents to take considerable interest in their child's progress in school and therefore it is arguably predictable that they will attach a higher importance to InCAS results than would be intended by the government. Where parents attach significant importance to an educational process, it follows that their children, their teachers and schools will attach similar if not higher levels of importance.

⁴ Figures relating to the CCEA surveys of pupils, parents, teachers and principals derive from the report: *Interim Findings, InCAS Evaluation 09/10*. All percentages are rounded and may not correspond exactly to sample and frequency counts. The questionnaire was sent to all schools for distribution to all of their parents in November 2009, i.e. after the occurrence and correction of the General Maths error. The deadline for return of the questionnaires was 22nd January 2010.

31. Pupils will be well aware that their parents will be attending the school to discuss the results of their InCAS assessments⁵ and in order to please them they will feel they must do their best. Whether they articulate it as such, they will feel they are being judged by their age-equivalent score; happy to be above it, unhappy to be below it. It is difficult, for pupils at least, to see the InCAS assessments as being simply diagnostic. While little evidence of pupil upset may be available it is reasonable to assume that being given one score on one occasion and being informed on another occasion that it was incorrect (and perhaps flattering) was to a greater or lesser extent distressing to those involved.

Parents

32. Not all of the parents of the pupils involved would have been re-briefed⁶ but it would not be unreasonable to suggest that those who did require repeat briefings would have been unhappy with the error situation. There is little doubt that they would have been greatly concerned and most probably the majority of them would have experienced disappointment when the discrepancies indicated that their children's age-equivalent scores should be downgraded. The concern would exist at two levels at least: concern about the impact on their children's motivation and concern about the extent to which they should attach importance to the scores now and in the future.

33. That said it is important to note that some 74% (11,146) of the 15,838 parents who responded to the CCEA survey felt that the InCAS results reflected what they considered to be their child's level of ability. Clearly 26% did not⁷ and the figure for the parents of children with special educational needs was even higher at 44% (representing 37 of the 94 who responded). These groups of parents may be reflecting a lack of confidence in the scores arising from the error situation or they might be over or under-estimating their child's level of ability in the contexts assessed. The former would be a particularly worrying matter but it is not possible from the survey data to discriminate between the two possibilities.

34. In the same survey, 85% and 82% of the parents of non-SEN pupils responded 'Yes' to a question seeking their views on whether the InCAS results helped to inform them of their child's strengths and areas for development respectively. The corresponding figures for parents of SEN pupils were 71% and 71% respectively. It would be unwise to invest too much importance in either set of responses, however, as they refer to quite broadly referenced strengths (i.e. relatively high age-equivalent scores) and weaknesses (i.e. relatively low age-equivalent scores) in General Maths and Reading rather than specific areas of numeracy or literacy. Even if the teachers used constituent scores to enhance the performance information for parents⁸, the areas remain very broad e.g. Measures, Shape and Space as a sub-scale of the General Maths score, and may not have much meaning for lay people.

⁵ It is a legal requirement that schools offer meetings to discuss InCAS results with parents: see the *Education (Assessment Arrangements) (Foundation to Key Stage 3) Order (Northern Ireland) 2007*

⁶ Clearly a sizeable proportion of the pupils' corrected scores would have differed little from the initial scores and a re-briefing could reasonably be deemed unnecessary. However, at least one of the schools, which responded to the Working Group's request for information, indicated that they had decided not to re-brief the parents choosing instead to discuss the matter in the following year. There was no indication how they addressed the fact that the Year 7 pupils would have left by the following year.

⁷ It is not possible to disaggregate missing and disagreement responses as the full analysis is not yet available.

⁸ In the case of General Maths this would entail going against the strong advice not to share sub-scale scores with parents

Teachers, Principals and Schools

Pressure on Schools

35. As mentioned above, the InCAS assessments are designed to be diagnostic and not high stakes. However, the importance of doing well in InCAS assessments is at the very least strongly suggested by evidence of commercially available practice tests. The extent to which these are used by teachers and schools is not known but uncorroborated comments from respondents to the Working Group suggest that it is a developing market.

It is recommended that research is undertaken into the extent of usage of commercial practice tests by schools and parents.

36. The reasons for schools attaching a high-stakes importance to InCAS would need some formal research but arguably can be traced to several influencing factors.
37. The first and most obvious would be the desire to ensure that pupils do their best they can in any assessment that is being reported externally, namely to parents. Just as obviously, there is the possibility of considerable pressure from parents that their child should be properly prepared for any assessment which compares their child to others in the class or year group/cohort etc. Emphasizing the diagnostic and non-normative approach to using the InCAS results is of little use when parents are given data in age-equivalent form; i.e. in a form that shows the child's score compared to the average for the specific age. It is therefore important to emphasize that the assessments are primarily for use by teachers in informing their teaching, with the key objective of improving the quality of the children's learning.

It is recommended that steps are taken to emphasize to parents that the primary purpose of the InCAS assessments is in contributing to diagnostic support for pupils' learning – and not for comparing pupils.

38. The second pressure on schools arises from two complementary demands, neither of which is intended to make explicit use of InCAS scores. These are the need to identify school weaknesses for addressing in school development plans and the other is the setting of targets for pupils' outcomes for the school to achieve. Both aspects are the subject of questions in the evaluation surveys for teachers and principals and the not-so-subtle effect of these twin demands is arguably for the teachers and principals (and by extension their pupils) to invest a higher priority in doing well in InCAS assessments rather than merely using the scores to assist with adjusting teaching to support individual pupils' needs. The trend towards making InCAS high stakes is likely to continue if it is not addressed.

It is recommended that research should be commissioned into how teachers and principals use InCAS scores for diagnostic, planning and evaluative purposes.

39. Information on the impact of the two errors on schools, their teachers and principals, was available to the chair of the Working Group in anonymized letters to CCEA and the Department of Education. There were surprisingly few of these with six letters or emails to DE, from two principals, one parent, a parents' group and two from teachers' unions; and five to CCEA from three principals, one teacher and a school governor. Five telephone calls were also logged by DE from four schools and one parent. Of the various inputs the principals' letters spoke most clearly to the incongruities in the results. Disappointment, puzzlement and in some cases anger are evident in the letters.
40. Additional information on impact was collected in two main ways: through direct consultation with teachers, parents and principals by members of the Working Group and by means of ad hoc surveys seeking written responses. The direct consultations

provided information on a variety of people's reactions to the errors. For example, it was reported that some teachers consider the compulsory assessment arrangements to be contradictory to the ethos of the Revised Curriculum. Some parents expressed confusion over the fact that testing at 11-plus is unacceptable but InCAS assessment is acceptable. In an Irish Medium education (IME) context, concern was raised about the ethics of using the assessments to trial new items that are not yet standardized or otherwise checked.

41. This is not just an IME matter, however, as the General Maths error for all pupils was caused specifically by the inclusion of trial items that were then included in determining age-equivalent scores⁹. As mentioned later, CCEA and CEM have confirmed that the location of the trial items at the end of assessments restricted any significant impact on the system's ability to judge each pupil's pathway through the sub-units and items. Put another way, it did not detrimentally impact on the choice of item which the system calculates 'on the fly' as having the appropriate level of difficulty to present to them next.
42. Placing the trial items at the end of an assessment is also considered to have restricted their impact largely to Year 7 scores but it remains the case that although CCEA were aware that the items were included, on the whole the teachers and pupils were not. Considered in the context of children being anxious about the assessments (reflected in a number of feedback responses), the inclusion of trial items is likely to be considered as an unfair extra burden on pupils when other means of trialling items are clearly available.

It is recommended that the trialling of items should be separate from formal use of the assessments, perhaps being developed in collaboration with schools that volunteer to participate.

It is recommended that if the formal use of the assessments must continue to incorporate trial items, this should be notified to pupils and teachers so that they know the items do not contribute to the assessment.

43. In the ad hoc survey, 20 responses were received from principals, broadly addressing the following four question areas:
 1. What direct experience of the errors did you have and what difficulties did they give rise to?
 2. Do you feel that the errors were dealt with appropriately? (Any discussion points in addition to 'yes/no' would be desirable)
 3. Now the errors are fixed, do you feel comfortable in using the the results you have?
 4. Do you feel the errors have had any lasting impact? If so, can you suggest ways in which confidence can be re-established where it has been lost?
44. The responses universally condemned the General Maths error and the compounding of its effect by the subsequent standardization error. Five schools, however, indicated that they had not experienced any difficulties arising directly from the errors, presumably because they were not among the 79 schools, according to CCEA data, that had already conducted parent meetings or the 83 that had already carried out the preparation for them.
45. The majority (14) experienced situations that variously included the need to re-run parent interviews, being unable to prepare the results for interviews that had already been scheduled (because the system was locked for a period while investigations were underway), and having to re-analyse and re-print outputs for all Years 4-7 teachers and each individual pupil. One response recorded a sense of humiliation at having conducted interviews with parents using flawed scores while another described the situation as

⁹ Trial items were included in the reading modules also but were not included in the determination of age-equivalent scores.

making their school look inefficient. Another described their colleagues being stressed through attempts to re-organize and re-interpret the corrected pupil data with only two days to go to the pre-scheduled interviews. One school reported having all but three of its pupils affected by corrected scores and having already completed all of their parent meetings.

46. Individual responses also included having to re-run two modules with their pupils, even though CCEA report that they had written and telephoned all affected schools to advise them that no action by schools was necessary. Another school decided not to pass on the corrections, citing the lateness of the notice of the error and the inadequacy of one day's substitution cost to cover the re-running of parent meetings. They intended discussing the matter in the next year but did not indicate how the Year 7 results (the set most affected by the error) would be addressed given that these pupils were leaving the school.

It is recommended that arrangements are made to test the performance of the InCAS data processing before general release of analysed age-equivalent and standardized scores. It would not be practical or sensible to do this for all schools as the response times for analysed data would probably be seriously extended. However, it could be accomplished by collaborating with a selection of volunteering schools which provide their data early in the autumn term and contribute to the necessary checking and feedback.

47. Of the 17 responses commenting on how the errors were dealt with, the respondents were split between those (9) who felt it was satisfactory (quickly corrected, apology appreciated, letter deflected parent tendency to blame the school etc) and those (8) who felt that not enough was done. In the latter category, respondents made a variety of points including the need to have stopped schools from proceeding with interviews immediately the error was known, and to have provided an explanation to schools more quickly (one principal reported knowing of the error but hearing the explanation on Radio Ulster). The point was made that the procedures that were undertaken did not change the fact that it was the teachers who had to cope with the problems, not the authorities. Several respondents, including those who felt that the matter was dealt with appropriately, commented that the subsequent support for substitute cover was inadequate.

It is recommended that the annual summer term briefings of relevant school staff on the upcoming autumn assessment period should include information on the measures taken to ensure the system is robust and dependable.

48. In response to the question asking whether they now felt comfortable, only four respondents agreed. One reported being initially comfortable but then having renewed doubts several months later after one pupil's scores were discovered to be incorrect even after the corrections were made. Most of the respondents (14) indicated a lack of confidence with the InCAS results, some wondering whether the problems were 'really fixed' and unlikely to re-occur, some comparing them poorly with other standardized and end of key stage test information, and some indicating that they perceived them to have limited diagnostic value in adjusting teaching approaches.

49. It cannot be deduced from these responses, or the CCEA survey data below, the extent to which teachers feel comfortable in their competence to share and explain InCAS results with parents. Oral feedback to the group did indicate, however, that some teachers have not had appropriate training and are not confident. For teachers such as these, the error in the General Maths only served to undermine what little confidence they had.

It is recommended that InCAS-related training should increase the emphasis on the interpretation and sharing of InCAS results with parents.

50. Five respondents felt that the errors would not have a lasting impact, either because the parents seemed not to be overly bothered or had accepted that errors can happen. That said, one of these respondents qualified their answer with the importance of ensuring a mechanism for identifying and correcting errors before the schools get the data, while another indicated that they would be more vigilant in the future. The majority (12) felt that there would be a lasting impact owing to a lack of trust among parents or because teachers continued to have reservations about the prevention of errors in InCAS results, the extent of their comparability with other standardized tests and their diagnostic potential.
51. Overall the sentiments expressed shock that the errors had been identified by schools; embarrassment as the errors reflected badly on the schools and the teachers; frustration at having to repeat time-consuming processes of data analysis and parent briefings; anger at having to carry out a compulsory task that was flawed beyond the control of the school; and a sense of vulnerability that it might happen again.

It is recommended that CCEA and DE take steps to ensure that the CEM quality control procedures are regularly and independently reviewed to reduce the potential for recurrences of errors of the types experienced in October 2009.

52. In any such unsystematic and largely opportunity-based survey¹⁰, the respondent group may be skewed to those who have relatively strong views that they wish to express. Though the responses reported above did indicate a variety of views, the CCEA survey would perhaps be expected to provide a wider perspective as it polled the whole population of schools including the 558 that had not been directly affected by the General Maths error. However, any consideration of the surveys involving principals and teachers needs to take into account that the subsequent standardization error compounded the fragility of the teachers' and principals' confidence in the system.
53. The interim results from the CCEA survey, carried out after the correction of results, show that 74% of the teacher respondents (345 of the 469 responding) felt that the INCAS feedback was consistent with their professional judgment. This is a considerable drop from the 94% (314 respondents) in CCEA's 2008/9 evaluation¹¹. The proportions of the remaining 26% disagreeing with the view or not responding to the item are not yet known. However, if the majority of these fell into the former category, with up to one quarter of the teachers not accepting that the InCAS scores reflected their own judgments, there is clearly cause for concern. The survey item is also not fine-grained enough to distinguish whether these teachers felt it is some or most of the pupils' results that were not consistent with their own judgments.
54. Of the teachers responding, 43 were teachers of special needs pupils and of those responding to the issue of consistency between InCAS results and their judgement, 61% (25 from 41 responses) agreed that it did. In this case the data does show unambiguously (i.e. excluding two non-responses) that 39% did not agree. A majority of respondents agreed that InCAS was manageable for their pupils (63% of responses, i.e. 26 out of the 38 responding to the item) but a slight majority of respondents (24) reported that their pupils did experience some difficulties.

Principals' Perspectives

55. The CCEA survey was completed by 182 principals of non-SEN schools. Their feedback demonstrates that they were largely satisfied with the overall manageability of the system (93%, 170 responses) but 46% (83 responses) reported that they had

¹⁰ Members were asked to canvass their own 'constituencies'

¹¹ *Final Evaluation Report on the Second Implementation Year of the InCAS Computer-based Assessments in 2008/2009*

experienced challenges in administering the system. These figures compare well with the previous year (when no errors were experienced) at 95% and 47% respectively. The interim report does not have information on these 2009/10 challenges but it would not be unreasonable to suggest that they also follow the previous year's pattern with comments on technical difficulties (logging on, setting up, computer crashes and freezing, problems with wireless access etc), cover for teachers administering the assessments, timetabling and computer access for large classes.

56. Particularly notable in the analysis of the principals' data is the fact that only 58% (105 respondents) felt that InCAS results reflected their pupils' levels of ability. With the caveat as before that the remaining 42% is not differentiated between those disagreeing with the view and those who did not complete the item, such a high percentage suggests there may be considerable disquiet among the responding principals.

It is recommended that meetings with principals are convened in the third term of 2009/10 to set out and give assurances on the steps being taken to secure the reliability of the system in the future.

57. Clearly principals cannot have direct experience of the correspondence of InCAS results with teacher judgments drawn from knowledge of the pupils' work over time and any additional information from other assessments. Instead this view would arise from staff discussion and consultation and this may have been influenced by direct experience of the original erroneous results and/or the much more widely felt concerns prompted by the media and public outputs from CCEA and DE. Data on principals' perceptions of the comparability are not available in the 2008/9 evaluation report so it is not possible to consider any possible impact from the error events of 2009/10.
58. Another change from 2008/9 is apparent in the 60% (110 respondents) who considered that the InCAS results helped to inform their school development plan. The corresponding 2008/9 figure was 50% (177 respondents). Both figures suggest, however, that sizeable minorities of up to 49% may not find InCAS useful in these processes. In 2009/10, 73% (133 respondents) also felt they helped in target setting but the interim findings do not offer information on how they are used in this context (no corresponding 2008/9 figure is available).
59. A small number (10) of principals of schools with SEN pupils responded to the survey. Six out of seven respondents agreed that the implementation was manageable but three indicated they had experienced some challenges in administering the assessments. Four of six respondents felt that the results did not reflect their pupils' levels of ability.

Diagnostic Information

60. Relatively high proportions of the teacher respondents in the CCEA survey, 81% and 79%, felt that the results were helpful in informing teaching and learning for individual pupils and their whole classes respectively. It should be noted, however, that these figures represent decreases from the 88% and 87% recorded for the same items in the 2008/9 evaluation and this may reflect an impact of the error events on the 2009/10 respondents' perceptions.
61. In contrast to the parents, teachers have the module and sub-scale scores of their pupils and the system bases its diagnostic credentials on the teachers being able to plan their teaching approaches on the basis of this finer-grained information. In the case of the General Maths error, however, only the overall age-equivalent score was corrected and for some pupils the sub-scale scores were higher than the overall score. This gave rise to puzzlement on the part of a principal who requested clarification of the reasons why one pupil's profile of sub-scale scores gave an arithmetic average that was higher than the final general maths score given. CEM and CCEA agreed that the school should be

informed that some items were used in the sub-scales which were not used for the General Maths scores.

62. This is a simplification of the case as the current InCAS system does not save item-level data for all pupils, retaining instead an age-equivalent score for each strand¹² of the assessments for some categories of pupils¹³. The correct analysis could be applied to the large majority of pupil scores but this problem of not having item-level data meant that another mechanism had to be found to provide corrected data for approximately 2,000 pupils that this affected.
63. A procedure to provide corrected scores from strand scores was tested using a different analysis system (Winsteps 3.64) and the known item-level data for 8,000 pupils. The corrected General Maths scores derived from the strand scores were demonstrated to match the scores based on the item-level data. CEM and CCEA subsequently agreed that the new General Maths age-equivalent scores for the pupils, for whom there were only strand scores, could be transformed from the original strand results without significant loss of accuracy.
64. CCEA and CEM also agreed that the sub-scale scores for all pupils would not be corrected, because all of them were considered to be within the normal confidence limits and were intended only for teacher use rather than for reporting to parents. This is arguably a reasonable prioritizing decision designed to correct the age-equivalent scores as quickly as possible for onward reporting to parents. Educationally, however, there must be reservations about potentially incorrect sub-scale diagnostics informing any adjustment of teaching; especially in cases where there was a large discrepancy between initial and corrected overall scores. It is also possible that the transformation processes and the lack of correction of sub-scale scores have fuelled some teachers' perceptions of the lack of diagnostic power of the results.

It is recommended that all schools should be asked to give specific feedback (perhaps through the annual evaluation surveys) on the diagnostic use and performance of the assessments.

Standardization Error

65. The design and intention of the InCAS requirements for schools does not include sharing standardized data from the assessments with parents. However, the reports considered by the Working Group suggested that schools wished to consider how pupils' performances within classes related to each other and the same year cohort. The system therefore offers standardized scores for all of the assessments and not just the Reading modules and General Maths.
66. The error that was identified a matter of weeks after the General Maths error was a failure to calculate correctly the standardized scores for all assessments. This resulted in exceptional scores being attributed to many more students than would be normal, mostly at the inflated end of the score range (116 or better). As with the General Maths error, the standardization error was easily rectified by the insertion of a line of code into the processing software, and after a short period of downtime the revised standardized scores were available to the schools. It is not possible to identify how many schools might have used the erroneous data to inform their teaching and learning, or their school development plan and target setting, but the data suggests that the maximum would be

¹² The assessments are presented in 'strands' that have items from across the 'sub-scales' i.e. the first strand of questions presented to a pupil may have items from the Number 1 and Measure, Shape & Space sub-scales.

¹³ The system does not currently save item-level data for a school once a threshold level of 30,000 items responses is reached for that school.

274. This is the number of 'hits' on the standardization data pages of the InCAS website but may include schools that visited the site more than once.

67. As mentioned the standardization data is only supplied for teachers' usage though there is nothing to stop teachers using the information to enhance discussions with parents. However, the real import of the error was less in its corruption of pupil-related data and more in its compounding of the effect of the General Maths error. Coming as it did after categorical assurances from CEM to CCEA, and onwards to DE and the schools, that no other errors existed, it had a devastating effect. The furore around the mistaken General Maths scores was re-ignited and confidence in the use of InCAS results plummeted.

CCEA and the Department

68. The use of InCAS is currently a compulsory requirement for schools with an overall investment of £2,606,544 since 2005/6 and a 2009/10 cost of £832, 674¹⁴ (to 31-12-2009). There is, therefore, a reasonable expectation from schools and the wider public that both organizations, from their separate perspectives and roles, should ensure that the system works efficiently and effectively. It almost goes without saying that staff at various levels in CCEA and the Department of Education have experienced considerable stress and additional workloads over both errors. The recent errors have created unnecessary difficulties ranging from addressing the formal questions of members of the assembly (including formal responses from the Minister) to creating a working group to look into the impact of the errors. Additional costs, amounting to £37,969 (figure calculated up to 31-12-2009) have been incurred for providing schools with one day's substitution costs (for each error event) to allow time for a member of staff to prepare the corrected materials and, where appropriate, meet with parents.

It is recommended that CCEA and DE seek to recoup the costs of substitution cover from CEM.

A Wider Perspective and Recommendations on Computer-based Diagnostic Assessment in Northern Ireland Primary Schools

69. This section considers aspects of the computer-based diagnostic assessment system that are wider than the specific error contexts of the previous section.

Compulsory Computer-based Diagnostic Assessment

70. It is not clear why schools are required to use a specific means of developing diagnostic information for their own use and for communication to parents. Prior to the 2007 legislation, schools used a variety of means to assess pupil progress and identify strengths and weaknesses. The legislation did not bar these various standardized and diagnostic tests and many schools have retained them, often using them to inform their judgment as a complement or alternative to the compulsory assessments. In its favour, legislating for a compulsory, specific means of diagnostic assessment should ensure consistency of approach and format of delivery. Over time it should also provide a series of measures that have a consistent basis for analysis and interpretation of progress. However, it is worth asking the question: "Are these reasons enough for it to be compulsory?"
71. A compulsory system has drawbacks that include the setting aside of significant core annual costs (whether at school, local authority or government department). The costings can also escalate when the choice of compulsory system is not subject to

¹⁴ Divided as £315, 756 and £516, 918 for software licensing and development/implementation costs respectively

competitive tendering either by the schools or local authorities/government departments. A compulsory system can also fall foul of a misunderstanding of its diagnostic intention and purposes, resulting, for example, in pupils, parents, teachers and schools attaching inappropriate importance to achieving better performances.

72. Statutory requirements to use computer-based adaptive diagnostic assessments are arguably unique to Northern Ireland and while this may not be a bad thing in itself, it does raise questions about why specific legislation has been enacted to require them to be undertaken.

It is recommended that the legislative dimension of computer-based assessments should be reviewed.

Item Banks

73. One of the problems that beset computer-based adaptive tests is that items can be over or under-exposed to pupil usage. Large item banks, usually of the order of 1,000 items upwards, are the best safeguard to memorization of questions and their answers caused by over-exposure. However, InCAS has several units that have very small numbers of items; for example the Reading modules have 79 items for word recognition, 60 for word decoding, only 10 passages for comprehension. Inevitably pupils will remember at least some of these words and passages from year to year while teachers who are so disposed will find it very easy to incorporate actual item words and parallel comprehension examples into InCAS practice sessions. A small number of respondents to the Working Group mentioned knowledge of the use of practice tests by teachers; 'dummy' runs through the assessments by teachers with their classes; and resits for pupils.
74. Despite the small sizes of some of the item banks, CEM's own standardization analysis¹⁵ suggests that under exposure is a feature, for example, in the Spelling assessment, which has only 92 items.
75. General Maths fares better with some 351 items but these are divided among four sub-scales (Number 1; Number 2; Measures, Shape & Space; and Data Handling). Years 4 and 5 will also have smaller numbers of items owing to the gradation in difficulty.
76. In addition to exposure-related problems, the relatively small numbers of items in sub-scales are arguably not able to provide good diagnostic information for teachers. This is recognized by CEM and the process that led to the General Maths error was intended to develop new items.

It is recommended that the sizes of item banks are considerably increased in order to improve diagnostic potential and to maintain credibility by reducing both the risk of memorization and the effectiveness of practising as a means to improving performance levels.

Differential Item Functioning (DIF)

77. A conventional concern in any major assessment systems is to ensure that items function without bias for all groups of pupils. At present there are no DIF analyses to show if the InCAS assessments perform differently between boys and girls, or with vulnerable groups of pupils such as those with English as a second language, or with children with special educational needs, or with children who have potentially disrupted schooling such as services children and traveller children. While the precise reasons are not known from the survey, the CCEA evaluation data indicate that almost 25% of the

¹⁵ *The Re-Standardization of InCAS (Northern Ireland) and the Links to Age*, January 27th, 2010

respondents with special educational needs (79 of the 330 responding SEN pupils) did encounter problems in using the system and a greater proportion (41%, 133 pupils) found aspects of the Reading assessments difficult.

78. Perhaps InCAS does not fall into a category needing an equality impact assessment but it would nevertheless be important to know whether InCAS functions differently with groups other than school Year groups.

It is recommended that differential item functioning analyses are carried out for sub-groupings within the pupil population including gender, second language, special needs, socio-economic status and minority ethnic groupings.

Standardized Scores

79. Responses from teachers, which indicate a lack of comparability of InCAS scores with other assessments, are to be expected and they do not necessarily reflect badly on InCAS. The extent to which such comparisons are possible with any degree of confidence is in itself problematic and would need to be established on a test-by-test basis. Clearly the ability to make comparisons is not part of the design or intentions of the InCAS system but teachers are familiar with age-equivalent or standardized scores from commercially available standardized tests and their availability from InCAS will continue to raise demands for comparability.

80. Such a comparison requires standardized scores but InCAS is unusual in the manner in which it creates these from age-equivalent scores. Standardized scores are normally derived from raw scores, i.e. the actual scores of the pupils in terms of marks for correct answers. Age-equivalent scores are already a derivation of raw scores before they are in turn transformed into standardized scores. Age-equivalence is not an interval scale in which equal measures on the scale have the same meaning, a basic requirement for a quantitative comparison of scores. For example, an InCAS 'measure' of one year below the pupil's age-equivalent score does not have the same meaning in developmental and progress terms for a 6-year old pupil and an 11 year old pupil. It is not clear why InCAS is designed in this manner.

It is recommended that InCAS consider offering standardized scores on the basis of raw scores and that these standardized scores should be referenced to both Year cohorts and age.

Standardization

81. Another aspect of standardization is the extent to which the existing item difficulties (originally created from a pilot sample in 2005/6) remain valid for one year to the next. It is clear that some items have become considerably 'easier', particularly in the Spelling module. It is also clear that owing to the relatively small numbers of children involved standardization of the Irish Medium assessments requires to be monitored every year. CEM have produced a thorough paper analysing the pros and cons of different time-frames for re-standardizing and this is under discussion with CCEA.

It is recommended that CCEA and CEM agree a policy on standardization as soon as possible.

82. There is also the perceived problem of repeat sittings in which, for example, pupils may improve their scores over a short space of time, 'dummy' tests used by teachers for demonstration or practice. It appears that data on the various types of inappropriate usage is currently difficult to obtain and the prevailing view is that such activities are not a significant problem. However, any form of 'cheating' (perhaps to improve scores to

assist with school development planning and target setting) has the potential to distort the validity of the standardization either by causing a change in item difficulty through practice and over –exposure, and clearly may not give a ‘true’ assessment of pupils who have practised for the assessments.

It is recommended that CEM institute appropriate modifications to ensure that information on the various types of inappropriate access to and usage of InCAS is available and traceable.

Hardware Accessibility

83. It is clear that accessibility for large groups and resources such as stable wireless broadband or hard-wired broadband facilities remain problematic for a sizeable minority of schools.

It is recommended that a review of resources for InCAS administration is carried out to identify schools needing support.

Appendix 1: Membership of Committee

Peter Archer (Educational Research Centre, Dublin)

Elisabeth Bingham (Cumran PS)

John Campbell (St Peter's PS)

Brian Currie (ETI)

John Gardner (Queen's University, Chair)

Avril Hall-Callaghan (NI Teachers' Council)

Ruth Kennedy (CCEA, Observer)

Arthur McGarrigle (NI Teachers' Council)

Pilib Mistéil (Bunscoil an tSléibhe Duibh)

Brian Morrow (DE, Observer)

Lexie Scott (Meeting 2, NI Teachers' Council)

Kevin Smyth (Meeting 1, NI Teachers' Council)

In Attendance

Declan Hyland/Julie Craig (DE)