Research Matters / 37
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

A short history of the Centre for Evaluation and Monitoring (CEM)

Chris Jellis (CEM)

To cite this article: Jellis, C. (2024). A short history of the Centre for Evaluation and Monitoring (CEM). *Research Matters: A Cambridge University Press & Assessment publication, 37*, 74–83. https://doi.org/10.17863/CAM.106033

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-37-a-short-history-of-the-centre-for-evaluation-and-monitoring-cem.pdf

Abstract:
The Centre for Evaluation and Monitoring (CEM), based in the North of England, recently celebrated its 40th birthday. Arising from an evaluation project at Newcastle University, and a subsequent move to Durham University, it rapidly grew in scope and influence, developing a series of highly regarded school assessments. For a relatively small organisation, its influence was seen across the world, resulting in outreach centres in New Zealand, Australia, and Hong Kong. Since being acquired by Cambridge University Press & Assessment in 2019, it has established itself in a unique role within the wider Cambridge organisation due mainly to its development of computer adaptive assessments for use in schools. This article documents the rise of CEM, from its early successes to its adoption of new ideas in educational assessment and supporting technology until the present day. However, CEM's development was not without its controversies, and these too make fascinating reading when set against the background of 40 years of ever changing educational policies.

A short history of the Centre for Evaluation and Monitoring (CEM)

Chris Jellis (Cambridge CEM)

The Centre for Evaluation and Monitoring (CEM), formerly the Curriculum, Evaluation and Management Centre (CEM) was acquired from the University of Durham in 2019 by a joint venture between Cambridge University Press and Cambridge Assessment. Since then, it has established itself in a unique role within the wider Cambridge organisation due mainly to its groundbreaking computer adaptive assessments for use in schools. What follows is not intended to be an exhaustive account of all the assessments created in the last 40 years of CEM, but more a focus on some of the highs (and lows) of major interest during that time. The history of CEM is an interesting one, emphasising as it does the crucial importance of diligent research and rigorous statistical analysis to back up the claims any assessment provider makes.

Beginnings

In 1981, Colin McCabe at Newcastle University won a contract to evaluate the Technical and Vocational Education Initiative (TVEI) in the North East of England. McCabe, with colleagues, established the Curriculum Evaluation and Management Centre to carry out this evaluation. TVEI was a government sponsored initiative designed to increase the uptake of work-related skills and qualifications. It was overseen by the Manpower Services Commission (MSC) to run in tandem with the newly created Youth Training Scheme (YTS) and gave rise to changes such as the establishment of BBC microcomputers in schools, along with the move to rebrand traditional subjects such as Woodwork and Metalwork as Design and Technology and Home Economics as Food Technology.

Among the staff of the newly formed CEM Centre was Dr Carol Taylor Fitz-Gibbon, a researcher and economist who had spent some of her early career in the USA and had an interest in demonstrating value in a fair way. In 1982, she was approached by a school governor who had a very simple question. The governor wanted to know whether the Mathematics A Level results from their school were good given their intake.

Carol realised that without equivalent data from other schools, the question could not reasonably be answered. She further realised that although A Level results had a strong effect on choice of profession and future career progression, very little research had been carried out in this area. To a researcher with a keen mind, it seemed an important question that needed answers. It became a significant feature of her later work.

COMBSE

In 1983, Carol established a research project named COMBSE (Confidential, Measurement Based, Self Evaluation) (Fitz-Gibbon, 1985) to find an answer to this intriguing question. The COMBSE project ran from 1983 to 1987. The plan was to collect O and A Level scores from local schools and pool the data to establish the link between the two examinations. She correctly predicted that the average O Level grade was the best indicator of each A Level grade and she also knew that the use of O Level results to predict likely A Level results could be a concern, because the O Levels were themselves the product of the schools. She therefore sought a measure of general ability and tried a number of standard high-level ability tests. None worked well, but she was able to use the International Test of Developed Abilities (ITDA) which was being developed under the auspices of the International Association for Educational Assessment (Fitz-Gibbon, 1996, p. 61). That worked as a good predictor when augmented with a vocabulary test.

In order to provide a more comprehensive picture of A Level success, students were asked about the ways in which they were taught and also invited to complete a questionnaire with closed and open questions about their feelings and attitudes. This comprehensive monitoring system produced a model for much of the subsequent monitoring projects developed at Newcastle and Durham.

COMBSE started with 12 schools agreeing to share their data, and when it came to an end it was being used by 47 schools. It was clear that as more schools contributed data to the project, the better was the outcome for all those involved. COMBSE had confined itself to reporting on A Level Maths and English results only. Could a new system be designed that could provide schools with information on a much wider range of subjects? It was now time to bring those skills and experiences gained from the TVEI evaluation and COMBSE together to create a wider reaching research project.

Alis

In 1989, Carol took over as Director of the CEM Centre and established a new school evaluation system to replace COMBSE. This new system was called the A Level Information System (Alis). In the same year Peter Tymms, a former teacher, and later to become Director of CEM, became the first Research Associate to work on the project.

The team were keen to build on the success of COMBSE, but it was clear that the use of O Level results to predict likely A Level results could be a concern, particularly as O Levels were imminently to be replaced by the General Certificate of Secondary Education, the GCSE. A measure of general ability that worked as a good predictor was therefore required. Some well-regarded assessments of general ability were tried, but none provided the predictive power required by the project. To this end it was decided that CEM should create their own bespoke measure of student ability. This new assessment, called the Test of Developed Abilities (TDA), proved to have a much greater predictive power

and became a standard part of the CEM testing model, not only for Alis, but for other CEM assessments that were to follow. The Alis system came to be used widely in secondary schools, providing as it did a measure of student ability and a prediction of future A Level results, which were vital for schools that were increasingly being measured by their outcomes.

Yellis

Following the success of the Alis system, considerations were made to create a similar system for younger students aged 14–16 that used predictions of likely GCSE results as the outcome rather than A Levels. This system, consisting of a new assessment providing a measure of general ability and a prediction of GCSE grades, was piloted in 1990 under the name Yellis (Year eleven information system). The pilot proved to be a success and the assessment was released to schools in 1992.

PIPS and ASPECTS

In the same year, an assessment for children in Year 6 of primary school was started. The new system was called PIPS (Performance Indicators in Primary Schools) and was soon modified to cover all year groups from Year 1 to Year 6. These were designed by the PIPS Director, Peter Tymms (Tymms 1999), who wrote the initial tests including the PIPS Baseline assessment for 4–5 year old children starting school in 1993. He also designed the feedback given to schools.

In 1994, Christine Merrell[1] was appointed, and her particular interest in Early Years education led to the development of an assessment for 3–4 year olds in nurseries, ASPECTS. PIPS Baseline was used by a quarter of primary schools in England in 1998, some schools having joined the project as part of a government initiative of national testing in the early years. It was replaced when the Early Years Foundation Stage (EYFS) was introduced in 2008, which involved a very different kind of approach to assessment (QCA, 2008).

The CEM approach to assessment

Carol's early work established some basic principles. Her goal was to use effective psychometric models that are good predictors of future achievement to create assessments that are dependable and fair. Another aim was to reduce the burden of assessment on teachers and students, which led rapidly to the adoption of computer adaptive testing. The main aim was to use the data from these assessments to provide teachers and school leaders with valid and reliable data upon which to make their decisions. Finally, there was the fundamental belief that teachers and leaders were in the best position to decide what to do with the data for their school.

1 Christine died recently after a short illness.

A new home

The organisation was growing and starting to have an influence upon school performance so, after some disagreements with Newcastle University in 1996, Carol was offered a new post at Durham University and moved the CEM Centre with her to the city of Durham, initially to offices close to the School of Education and then to larger premises on the Durham University Science Campus. Along with this success came more money, allowing the establishment of new posts, and among those appointed at this time were Robert Coe and Kate Bailey, both to become future directors of CEM. Also at this time, a pilot for a new assessment for students aged 11–14 in secondary school was launched. This assessment, known as MidYIS (Middle Years Information System) provided a measure of student ability, plus a prediction to GCSE.

InCAS

In 2002, CEM launched InCAS (Interactive Computerised Assessment System), a groundbreaking new computer adaptive assessment which used a single piece of software to cover the age range from 5 to 11. Data from each of the PIPS assessments in Years 1 to 6 (ages 5 to 11) were analysed using the Rasch statistical method, enabling the team to establish a single scale in each of the key cognitive areas for the whole primary range. These scales were then used to build a single computer adaptive test. Students would start the assessment with items easy for their age and through adaptive testing their ability level would be established and recorded as an age equivalent score. The system provided a reliable and efficient way of measuring student abilities. As students took the assessment each year, a measure of longitudinal progress of their time in primary school was established. InCAS went on to be adopted for a number of years as a mandatory assessment for use in state primary schools in Northern Ireland.

BASE

In 2015, the UK government planned to mandate a baseline assessment in the reception classes of English state schools, to provide teachers with a measure of what pupils knew and could do when they started school. CEM had been running the PIPS baseline assessment successfully for many years, and grasped the opportunity to develop a new baseline assessment along the general lines of PIPS but updated to take into account the feedback received from teachers and schools over this time. CEM was now under the directorship of Robert Coe and CEM's bid was successful. The subsequent assessment, known as BASE, became one of the mandated reception assessments for the next two years. After this time, government policy changed under pressure from unions and other lobbyists and mandated reception assessment was dropped (in 2021 it was reintroduced in yet another form). The BASE assessment, however, continues and is used around the world.

iPIPS

Although not a CEM commercial product, the iPIPS system was developed by Peter Tymms to provide information for policy makers about what is happening in the

first year at school. It involved translating the original PIPS baseline assessment into many different languages The iPIPS system has been used to great effect in Brazil, Lesotho, South Africa and Russia, and the findings from those studies form the subject of a book (Tymms et al., 2023).

Check Together

The first assessment produced after CEM joined Cambridge in 2019 was a modified version of the BASE assessment specifically designed for use in Cambridge schools in India. This version, featuring a uniquely Indian soundtrack, imagery, content, and reports was developed in collaboration with colleagues in Cambridge.

The Cambridge Wellbeing Check

Following the Covid-19 pandemic in 2020, and the detrimental effects caused to school pupils due to school closures, greater emphasis started to be placed on student wellbeing than had previously been the case. The Cambridge Wellbeing Check was developed from a survey developed by researchers Dr Ros McLellan, Maurice Galton, Susan Steward and Charlotte Page in the University of Cambridge's Faculty of Education (McLellan & Steward, 2015). The original survey was created as part of a study examining the role of creative initiatives in fostering wellbeing, which was funded by the international creative learning foundation Creativity, Culture and Education. CEM has since worked with Dr McLellan and her colleagues to refine the questionnaire. It is now administered as a digital check for students aged 7 and above, alongside materials teachers can use to support school wellbeing initiatives.

Preliminary work is now being carried out to further integrate wellbeing with other CEM assessments and provide greater insights.

Cambridge Early Years Check Together

Following the development of Check Together in India, Cambridge colleagues requested a version of the assessment to augment the newly developed Cambridge Early Years curriculum. A new soundtrack, graphics and content were developed with the view to provide an assessment appropriate for as wide an audience as possible, along with greater integration with the Cambridge Early Years curriculum. The assessment was launched in the autumn of 2023.

Controversy

CEM's story has been intertwined with the Department for Education (DfE) and their initiatives for a long time, providing both support and challenge. Although originally established to evaluate the Technical and Vocational Education Initiative (TVEI), that evaluation and subsequent report (Fitz-Gibbon et al, 1988) found worse outcomes for those students that had been involved in the TVEI project than those that had not. Considering that the TVEI project had a budget of £900 million, this was quite a blow and was not received well.

Similarly, in 1999 the Education Secretary David Blunkett, hit out at CEM researchers (TES, 1999) who challenged the government view that older primary school children should be set 30 minutes of homework each night. CEM's research involving a survey of 20 000 pupils found that those who were set homework just once a month achieved better test scores.

In 2001, Professors Tymms and Fitz-Gibbon (2001) challenged the validity and accuracy of government figures regarding the increase in standards of Key Stage 2 results. Their work examined exam results over the previous 25 years and found some rise in standards, but not to the extent claimed by the government.

Again in 2004, Professor Tymms published an article in the *British Educational Research Journal* (Tymms, 2004) questioning the government's claims that literacy standards among 11-year-olds had risen dramatically between 1995 and 2000. This enraged the then Education Secretary, Ruth Kelly (Mansell, 2005), but Tymms' central argument was backed by the Statistics Commission, a non-departmental public body set up to oversee the work of the Office for National Statistics which refused to change its view, even in the light of heavy government pressure. The Statistics Commission's report (Statistics Commission, 2005) included a letter from Tim Oates, then head of research and statistics at the Qualifications and Curriculum Authority (QCA) which also supported Tymms' position.

Carol Fitz-Gibbon and Peter Tymms also came under pressure from statisticians to use multilevel models when analysing school data. In fact, Carol had considered using multilevel methods early in the development of the Alis assessment and wrote a paper discussing the use of such models (Fitz-Gibbon, 1991). Although acknowledging the strengths of the method, she ultimately rejected it for use in the Alis system as she felt that using a simpler system would be easier to explain to school personnel. Nevertheless, Carol Fitz-Gibbon and Peter Tymms were invited to explain their approach in a meeting at the Department for Education with Harvey Goldstein (a member of the Royal Statistical Society and a leading proponent of multilevel modelling), and Nick Tate (chief curriculum and qualifications adviser to the Secretary of State for Education). They were able to successfully argue their case.

> "Harvey said 'you've got to use multilevel models' and in fact we said 'no, no, no. If you look at the results in multilevel models, they are exactly the same as the ones you get out of classical tests' and we had a meeting, a showdown with Harvey at the DfE under Nick Tate and we won the argument against Harvey. I don't think we were ever forgiven for that." (Peter Tymms, personal communication)

Carol and Peter's work with Luke Vincent on the comparative difficulty of A Level subjects (Fitz-Gibbon and Vincent, 1994; Tymms and Vincent, 1995) resulted in further criticism from Harvey Goldstein and Michael Cresswell (Goldstein and Cresswell, 1996), this time focusing on their use of the subject pairs analysis approach and the use of Alis data in the analysis. The controversy continued for

some time and was addressed again in 2008 (Coe et al., 2008) by a team led by Robert Coe, who went on to become the Director of CEM in 2010.

New ideas

Carol Fitz-Gibbon had previously worked in the USA and brought some of the prevailing ideas about education measurement with her when she returned to the UK. One of these ideas was the concept of value added. Following the success of Alis, she won a contract to set up a value-added system in Scotland for Highers using Standard Grade results as the baseline, which lasted for many years (Fitz-Gibbon, 1992).

This piqued the interest of the Westminster government. In 1995 it commissioned a contract to research a new model for measurement of school outcomes. CEM won the contract, and in 1997 the Value-Added National Project report was published (Fitz-Gibbon, 1997). The report recommended a method of determining value added and a variation on the general approach was then adopted by the government.

Carol was also a great advocate of the Randomised Control Trial (RCT) (where subjects are randomly assigned to one of two groups, experimental and control) and was influential in the creation of the Campbell Collaboration project in the USA. At the time it was extremely unusual to use RCTs in educational research but subsequently they were used to great effect by CEM staff in peer learning projects in Scotland (Tymms et al., 2011). It was the first randomised control trial for peer tutoring that went across a whole local education authority, and it is believed to have been the largest randomised control trial in education at the time. Now RCTs are widely used in education.

For many years in the UK, analysis of test results from examinations and other assessments used a model called Classical Test Theory (CTT). Carol realised that a newer model, called Item Response Theory (IRT) was being used extensively in other countries, particularly the USA and Australia. She advocated its use in the UK too but fell foul of some of the leading statisticians in the UK, who felt that the model was not appropriate (see for example Goldstein, 1979; Panayides et al., 2010). Undaunted, Carol continued and the IRT model is now used extensively in CEM assessments. To establish greater interest in IRT measurement, Peter Tymms held a meeting at Durham University of likeminded people who were working with the Rasch model, including Tom Bramley from Cambridge Assessment. This established the UK Rasch User Group, which has met regularly for many years, and of which Cambridge is a very active member.

Outreach

From its earliest times, CEM has had an effect, not only on education in the UK, but also around the world. In 1998, CEM established a relationship with the University of Western Australia and established a CEM outreach centre there with Helen Wildy as director. A year later CEM established another outreach centre in New

Zealand, followed by one in Hong Kong in 2001. These centres were able to foster regional interest in CEM assessments and research and reach a much greater audience than could be achieved from the UK alone.

Research

As CEM expanded, its research section grew accordingly. The section rapidly gained attention as a centre for excellence and won many contracts from organisations such as the Sutton Trust and the Education Endowment Foundation, contributing significantly to Durham University's research excellence framework (REF) submission. Many studies, such as the peer learning study in Fife, Scotland (Tymms et al., 2011) and various explorations into the nature of ADHD (Attention Deficit Hyperactivity Disorder) manifestation in the classroom (Sayal et al., 2020), have also used CEM assessments as pre- and post-measures of ability when investigating potential educational interventions.

Present day

Currently, all CEM assessments are delivered digitally, and work has been ongoing to explore how the capabilities in CEM can be brought to bear on enhancing the Cambridge offer to schools in the UK and overseas. Kate Bailey, who started in CEM in 1996, is now the Managing Director, replacing Elizabeth Cater who headed CEM after the integration with Cambridge. Current and previous CEM directors have recently published a book, *The First Year at School: An International Perspective* (Tymms et al., 2023), which details work on the iPIPS project and its effect around the world. The book is dedicated to Christine Merrell who created the PIPS and ASPECTS baseline assessments with Peter Tymms and created the original design for the BASE assessment.

CEM's focus for the future will be on strengthening the baseline assessments that CEM is known for and ensuring that they can support all Cambridge schools in improving the outcomes of learners all round the world. There is more to do in exploring how the unique capabilities in CEM can be used to enhance the Cambridge portfolio and reach even more learners in future.

References

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects*. CEM centre.

Fitz-Gibbon, C. T. (1985). A-level Results in Comprehensive Schools: the COMBSE project, Year 1. *Oxford Review of Education, 11*(1), 43–58.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In *Schools, classrooms, and pupils* (pp. 67–83). Academic Press.

Fitz-Gibbon, C. T. (1992). *Performance indicators and examination results.* Scottish Office Education Department, Research and Intelligence Unit.

Taylor Fitz-Gibbon, C. (1996). Monitoring Education. *Monitoring Education: Indicators, Quality and Effectiveness*. Bloomsbury.

Fitz-Gibbon, C. T. (1997). *Feasibility studies for a national system of Value Added indicators*. SCAA.

Fitz-Gibbon, C. T., & Vincent, L. S. (1994). *Candidates' performance in science and mathematics at A-level*. School Curriculum and Assessment Authority.

Fitz-Gibbon, C. T., Hazelwood, R. D., Tymms, P. B., & McCabe, J. J. C. (1988). Performance indicators and the TVEI pilot. *Evaluation & Research in Education, 2*(2), 49–60.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5*(2), 211–220.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education, 22*(4), 435–442.

Mansell, W. (2005, August 26). *Why this man scares Ruth Kelly*. TES.

McLellan, R., & Steward, S. (2015). Measuring children and young people's wellbeing in the school context. *Cambridge Journal of Education, 45*(3), 307–332.

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal, 36*(4), 611–626.

QCA. (2008). *Early years foundation stage Profile handbook*. Qualifications and Curriculum Authority.

Sayal, K., Merrell, C., Tymms, P., & Kasim, A. (2020). Academic outcomes following a school-based RCT for ADHD: 6-year follow-up. *Journal of Attention Disorders, 24*(1), 66–72.

Statistics Commission. (2005). *Measuring standards in English primary schools*. Statistics Commission Report, 10.

TES. (1999, July 23). Too heavy with elite (p. 12).

Tymms, P. (1999). *Baseline assessment and monitoring in primary schools*. David Fulton Publishers.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal, 30*(4), 477–494.

Tymms, P., & Fitz-Gibbon, C. (2001). Standards, achievement and educational performance: a cause for celebration? In R. Phillips, & J. Furlong (Eds.), *Education, reform and the state: Twenty-five years of politics, policy and practice*. Routledge.

Tymms, P. B., & Vincent, L. (1995). *Comparing Examination Boards and Syllabuses at A-level: Students' Grades, Attitudes and Perceptions of Classroom Processes: Technical Report*. Northern Ireland Council for the Curriculum, Examinations and Assessment.

Tymms, P., Bartholo, T. Howie, S. J., Kardanova, E., Koslinski, M. C., Merrell, C., & Wildy, H. (Eds.). (2023). The first year at school: An international perspective, *International Perspectives on Early Childhood and Development, 39*, Springer.

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K., & Miller, D. (2011). Improving attainment across a whole district: School reform through peer tutoring in a randomized controlled trial. *School Effectiveness and School Improvement, 22*(3), 265–289.