Assessment in Senior Secondary Physical Education. Questions of judgement

Dawn Penney[1], Lorna Gillespie[2], Andrew Jones[3], Paul Newhouse[3], and Alistair Campbell[3]

1. *Faculty of Education, University of Waikato, New Zealand and School of Education, Edith Cowan University, Australia*
2. *Faculty of Education, University of Waikato, New Zealand*
3. *School of Education, Edith Cowan University, Australia*

The ways in which various aspects of senior physical education courses should be assessed and whether some can, or indeed should be incorporated in external examinations, are matters of longstanding professional debate across Australia and internationally. Differences in current practice across Australasia reflect an ongoing lack of consensus about how assessment requirements and arrangements and particularly, examinations in senior physical education, can best address concerns to ensure validity, reliability, equity and feasibility. An issue never far from such debates is that of 'professional judgement' and more specifically, whether and how professional judgement does and/or should 'come into play' in assessment. This paper reports on research that has explored new approaches to examination assessment and marking in senior physical education, using digital technologies. It focuses specifically on the ways in which 'professional judgement' can be deemed to be inherent to two contrasting methods of assessment used in the project: 'analytical standards-based' assessment and 'comparative pairs' assessment. Details of each method of assessment are presented. Data arising directly from assessors' comments and from analysis which explored inter-marker reliability for each method of assessment and compared results generated by internal teacher assessment, standards-based and comparative pairs assessment, is reported. Discussion explores whether the data arising can be seen as lending weight to arguments for (i) more faith to be placed in professional judgement and (ii) for the comparative pairs methods to be more widely employed in examination assessment in senior physical education.

Introduction

Assessment remains arguably one of the most contentious issues in relation to senior secondary physical education. The ways in which various aspects of senior physical education courses should be assessed and whether some can, or indeed should be incorporated in external examinations, are

matters of longstanding professional debate across Australia and internationally. In many instances, these debates have centred on physical education as a 'performance-based subject' (Macdonald & Brooker, 1997) and more specifically, the extent to which assessment requirements have supported efforts to promote interconnections between 'theoretical' and 'practical' dimensions of learning (see for example, Kirk, Penney, Burgess-Limerick, Gorely, & Maynard, 2002; Kirk, Burgess-Limerick, Kiss, Lahey, & Penney, 2004; Macdonald & Brooker, 1997; Thorburn 2007). Differences in assessment requirements and arrangements in senior physical education courses internationally reflect ongoing dilemmas and challenges in this regard. They also reflect a lack of consensus about how requirements and arrangements relating to assessment and particularly, examinations in senior physical education that are linked to tertiary entrance, can best address concerns to ensure validity, reliability, equity and feasibility.

As many readers will be aware, across Australia some senior physical education tertiary entrance courses feature a mandatory external examination, which in some instances incorporates a practical component. The respective weighting that is accorded to the external examination component and school-based assessment also varies. In Queensland standards-based teacher assessment is undertaken for all senior secondary authority subjects and there is no external examination component. In New Zealand NCEA (National Certificate of Educational Achievement) Physical Education (Level 1, 2 and 3) similarly has only internally assessed standards, meaning that all assessment is school based, carried out by teachers. Scholarship Physical Education (designed for the very best students and awarded to approximately 3% of the Level 3 cohort nationally) is examined solely via an external written examination (New Zealand Qualifications Authority, 2010). In Western Australia (WA), (the context of our research), students who are studying at least one pair of units from Stage 2 or 3 of the Physical Education Studies (PES) course sit a written and a practical (performance) examination in their final year of study (unless they are exempt). Stage 2 and 3 units represent the highest level of study in the course and each unit usually represents a semester of work. The examination comprises a written examination worth 70% of the total examination score and a practical (performance) examination worth 30% of the total examination score (Curriculum Council of WA, 2010).

A matter never far from debates about the merits of particular course and system requirements for assessment is that of 'professional judgement' and more specifically, whether and how professional judgement does and/or should 'come into play' in assessment. In raising this issue, we deliberately distinguish it from 'teacher judgement' and in so doing, seek to avoid attention centring on the merits or shortcomings of teacher as compared to external judgement. Our stance is that 'professional judgement' is inherent in all assessment. Our interest is in the particular nature of the

judgement and the processes associated with making judgements, in the context of two contrasting methods of assessment used in a research project undertaken in WA.

The digital assessment project

The three year project has been supported by funding from the Australian Research Council (ARC) and Curriculum Council of Western Australia. It has utilised digital technologies in exploring new approaches to examination assessment in *Physical Education Studies* and three other senior secondary courses; *Engineering Studies, Applied Information Technology Studies,* and *Italian.* The key point of commonality in selection of these courses was a concern for assessment in an examination context to directly engage with performance in authentic tasks. In referring to 'authenticity' we fore ground Thorburn's (2007) emphasis of the need for assessment in senior physical education that aligns with the conceptual underpinnings of contemporary curriculum documents, and Hay's (2006, p. 317) view that authentic assessment 'should be based in movement and capture the cognitive and psychomotor processes involved in the competent performance of physical activities'.

Details of the assessment tasks developed for the project, specific information about the new senior physical eductaion studies course in WA and initial data from teachers and students involved in the project, have been provided elsewhere (Jones, Penney, Newhouse, & Campbell, 2009; Penney, Newhouse, Jones, & Campbell, in press). Here we provide an outline of the task framework before directing attention to the two methods used to assess student performance in the tasks: *'analytical standards-based'* assessment and *'comparative pairs'* assessment.

As explained previously (see Jones et al., 2009) the project sought to generate various forms of digitally based representations of student performance in the assessment tasks, enabling reliable assessment for the purposes of external examination to be explored. Each component of an integrated task generated digital evidence as follows (for further task and technical details, see Jones et al., 2009):

- Part 1. Structured on-line response to a tactical problem in a specific activity context: *text and graphic format responses*;
- Part 2. Performance of four skills pertinent to the tactical problem: *video recordings of student performance;*
- Part 3. Application of skills in a game/competitive performance context: *video recordings of student performance;*
- Part 4. Structured on-line reflection on performance: *text and graphic format responses*.

The digital representations of student work generated from each part were uploaded to an online repository to be accessed by assessors, with all evidence for each student collated into a unique record. All marking was undertaken on-line and the same digital evidence was used in both methods of assessment.

For the *analytical standards-based assessment*, standards-based rubrics were developed for each part of the examination task with the PES course document used as the reference point in development (see Jones et al., 2009 for a complete task rubric). The on-line marking interface enabled the rubric and student evidence to be displayed simultaneously via a split screen. Assessors were required to make a number of judgments in relation to evidence from each part of the task. In each instance this involved a decision directly linked to the rubric. For example, one of the judgments required for 'execution of the strategic response in a 'live' performance context (modified game situation) (part 3) focused on students' ability to 'make on-the-spot decisions to apply movement patterns in solving tactical problems', with the rubric describing five standards. A judgment required in both part 1 and part 4 focused on students' understanding of the tactical concepts of games and activities, again with the five standards described in the rubric. The on-line marking system enabled assessors to choose the order in which they undertook judgments relating to each part of the task and return to any judgment or aspect of student evidence at any time. For the analytical standards-based assessment, two assessors were assigned for each student's work.

Comparative pairs assessment was then undertaken by a number of assessors (five in years 1 and 2 and 20 in year 3). As the name suggests, this method requires a direct comparison of two students' performance in the task, with each student's full evidence record from all parts of the task available to assessors. Pairings are automatically generated and assessors are required to decide which of the two students being compared has performed better in relation to the set criteria. A further point to note was that the task was adapted for various activity contexts and that comparative pairs marking therefore included comparisons of performance in different activity contexts. We acknowledge that task comparability and assessment across contexts are both matters worthy of extended discussion. While that is beyond the scope of this paper, the inclusion of varied activity contexts is important to note in considering the nature of judgments being made and processes involved in reaching a judgment.

Two different systems for comparative pairs marking have been used during the project. A brief description of each is provided here because of our interest in exploring the judgment process. Having been involved in the project as both researchers and assessors, we suggest that differences in the comparative pairs marking interface may mean that there are also notable differences in the process assessors go through to reach a judgment.

In years 1 and 2, the comparative pairs marking interface enabled a split screen simultaneous side-by-side presentation of the two students' work. Assessors could access and compare any of the students' digital files from parts 1-4 of the task. They were required to make a total of four judgments for each pairing: three judgments relating to specific assessment criteria and one overall, holistic judgment. In each instance, the judgment required was a simple decision: *Is A better than B or vice versa?* Accompanying guidance for each criterion indicated the evidence deemed most pertinent to the decision required. For example, it was stated that evidence relating to criterion 2 'Execution of movement skills' would be drawn from the individual skill performance extracts and may also be drawn from game play/ competitive performance video extracts. In the second year, a field was added to enable markers to record their comments on the students' work, so that they would not have to scroll through pages or view complete videos of the student work each time a particular student appeared in a pairing.

In the final year of the project a different comparative pairs interface was used. A key advantage of the new system was that reliability scores are generated on an ongoing basis, such that marking can cease once an agreed reliability score is achieved (rather than reliability scores only being able to be calculated after all marking has been undertaken). In undertaking comparative pairs assessment using this interface, assessors were only required to make a single judgment, *'who wins – A or B?'* based on a holistic criterion. All digital evidence for both students in the pair was available and assessors could switch between students at any time and make notes about aspects of evidence, but could not simultaneously view evidence from both students.

Comparative Pairs Assessment: Key findings

Analysis of assessment data generated during each year of the project has explored inter-marker reliability for each method of assessment and has compared results generated by internal teacher assessment, standards-based and comparative pairs assessment. Data has also been gathered directly from assessors via interviews, with questions addressing both the task and the marking process. Below we present some key findings arising from the assessment data that are pertinent to consideration particularly, of whether there is a case for the comparative pairs method to be widely employed in examination assessment in senior physical education.

In year 1 data was gathered from four classes of Year 11 students (n=39), each focusing on a different activity context. Comparative pairs assessment involved five assessors making a total of 745 judgements. Separation Indexes (SIs) were calculated for each of the three specific criteria and the holistic criteria, to provide an indicator the overall internal consistency of judges. Sis are given as a

number from 0 to 1 with values closer to 1.00 being more desirable. This generated a reliable set of scores (SIs of 0.905 to 0.929) with scores and rankings highly correlated to that of the two external assessors using the analytical standards-referenced method ($r = 0.88$). The teachers' marks for the assessment item were significantly moderately correlated with the results from the comparative pairs marking ($p<0.01$).

In year 2 data was again drawn from four year 11 classes all featuring different activity contexts and involved five assessors making a total of 250 judgments. Assessment of 27 students' work using the comparative pairs method again produced a reliable set of scores ($SI = 0.75$) that was significantly correlated to the analytical marking scores ($r = 0.69$, $p<0.01$). The teachers' marks for the assessment task were significantly moderately correlated with the results from the comparative pairs marking ($r= 0.54$, $p<0.05$). This was also true for the rankings where the rankings of the teachers' marks were correlated with the comparative pairs method of marking (statistically significant, $p<0.05$) but not analytical.

In the final year of the project, 20 assessors were involved in comparative pairs assessment of the exam output for 108 students from 11 different year 11 class groups utilising six different activity contexts. It had been decided to stop marking once the Cronbach Alpha Reliability Coefficient was above 0.95. This occurred after the 13th round of marking when a total of 710 judgements had been made, with a coefficient of 0.958. The system also provided statistics on the consistency of the 20 judges, which showed that 46 (6.5%) of the 710 judgements appeared to be seriously inconsistent. There was a significant moderate correlation ($r=0.73$ $p<0.01$) between the scores generated by comparative pairs marking and the score determined by analytical marking, but also some notable differences in scores and rankings generated by the two methods. The ranking from the individual analytical assessors tended to be correlated low to moderate with the ranking from the pairs judging ($r=0.21$, $p<0.05$ and $r=0.62$, $p<0.1$). For some students there were substantial differences in the scores awarded by the different methods of marking and in the overall ranking in the population. Further investigation of data is pursuing these differences. There was a moderate to low significant correlation between the teacher's score and the pairs judging score ($r=0.46$, $p<0.01$) and the teacher's semester mark ($r=0.39$, $p<0.01$).

Discussion

The project has explored the application of the comparative pairs method of assessment in the context of a complex task in senior physical education studies. In each year a reliable set of scores have been generated via this method, that have also been shown to reasonably align with the results of analytic marking and teacher assessment. Some differences arising from the different methods require further

investigation. The findings provide a clear case for further research investigating the use of the comparative pairs method of assessment in physical education.

Feedback from assessors involved in the final year indicated their ease in working with the marking system, and pointed to recognition of some difficulty in making judgements between students undertaking the task in different activity contexts and/or when two students' performance in the task was deemed to be very similar. Certainly, the task itself *and* the comparative pairs method are both influential in terms of the type of judgement required and processes involved in reaching a judgement. The differences in the comparative pairs marking interfaces in years 1 and 2 as compared to in year 3, can also in some ways be seen as having potentially important implications for the judgement process – and potentially, therefore, the judgement made.

The data and our experiences of making judgements using the comparative pairs method have raised questions for us about both the *nature* of professional judgement and *processes* of professional judgement inherent in the comparative pairs method, and how this may differ from judgements made using the analytic method. From our personal experience we relate what was required of assessors using the comparative pairs method to judge students' performances in this task, to the notion of complex judgements as 'configurational assessment' (Kaplan, 1973, as cited in Crisp, 2010). As Crisp (2010, p.3) explains, the suggestion is that 'judges have compounded criteria onto one uni-dimensional scale, not via a set of rules but through mental integration, and they unpack some of the criteria from this scale when justifying their judgements'. This and other issues arising from our data and project experience are undoubtedly worthy of professional discussion and further exploration through research.

Acknowledgement

The research discussed in this paper is as a result of the work of a research team organized by the Centre for Schooling and Learning Technologies at Edith Cowan University (http://csalt.education.ecu.edu.au/). The team was led by Paul Newhouse and John Williams and included senior researchers Dawn Penney, Cher Ping Lim, Jeremy Pagram, Andrew Jones, Martin Cooper, Alistair Campbell, project managers and many research assistants. The work of everyone in this team has contributed to the research outcomes presented in this paper.

References

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*(1), 1-21.

Curriculum Council of Western Australia (2010). *Physical Education Studies*. Perth, Australia: author.

Hay, P. J. (2006). Assessment for Learning in Physical Education, in D. Kirk, D. Macdonald, & M. O'Sullivan (Eds.), *The Handbook of Physical Education* (pp. 312–25). London: Sage.

Jones, A., Penney, D., Newhouse, P. & Campbell, A. (2009). Digital assessment in high stakes Physical Education practical examinations. In T. F. Cuddihy & E. Brymer (Eds.), *Creating Active Futures. Edited Proceedings of the 26th ACHPER International Conference* (pp.217-232). Brisbane, Australia: Queensland University of Technology.

Kirk, D., Penney, D., Burgess-Limerick, R., Gorely, P., & Maynard, C. A. (2002). *The Reflective Performer in Physical Education: A Complete Guide to A-Level Study*. Champaign, IL: Human Kinetics.

Kirk, D., Burgess-Limerick, R., Kiss, M., Lahey, J., & Penney, D. (2004). *Senior Physical Education. An Integrated Approach* (2nd ed.). Champaign, IL: Human Kinetics.

Macdonald, D., & Brooker, R. (1997). Assessment issues in a performance-based subject : A case study of physical education. *Studies in Educational Evaluation, 23*(1), 83-102.

New Zealand Qualifications Authority. (2010). *Scholarship.* Retrieved from http://www.nzqa.govt.nz/qualifications-standards/awards/scholarship/

Penney, D., Newhouse, P., Jones, A. & Campbell, A. (in press). Digital technologies: Enhancing pedagogy and extending opportunities for learning in senior secondary physical education? In T.Le & Q.Le (Eds.). *Technologies for Enhancing Pedagogy, Engagement and Empowerment in Education: Creating Learning-Friendly Environments*. Hershey, P.A., USA: IGI Global.

Thorburn, M. (2007). Achieving conceptual and curriculum coherence in high-stakes school examinations in Physical Education *Physical Education and Sport Pedagogy, 12*(2), 163-184.