RESEARCH REPORT

(Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject

Peter J. Hay* and Doune Macdonald

School of Human Movement Studies, University of Queensland, St. Lucia, Queensland, Australia

This paper draws on semi-structured interview data and participant observations of senior secondary Physical Education (PE) teachers and students at two school sites across 20 weeks of the school year. The data indicated that the teachers in this study made progressive judgements about students’ level of achievement across each unit of work without explicit or overt reference to the criteria and standards represented in the schools’ work programmes and in the Senior PE syllabus. The teachers’ justification for such an approach was that the criteria and standards had become for them sufficiently ‘internalised’. Determining students’ levels of achievement was for the teachers somewhat ‘intuitive’, being reliant on their memory of students’ performances, and influenced by the construct-irrelevant affective characteristics of the students. It is argued in this paper that such construct-irrelevant compromised the construct validity and possible inter-rater reliability of the decisions made and advantaged some students and marginalised others on the basis of characteristics that were not specifically related to the learning expected from following the syllabus. The potential inequities of such an approach are discussed and suggestions are made for the consolidation of the validity and reliability of teachers’ judgements.

Introduction

Rink and Mitchell (2002, 209) suggested that ‘[o]ne unintended outcome of the standards, assessment and accountability movement is that any program not included in high-stakes state level assessment, for all practical purposes, does not “count”’. An example of a PE programme that engages students in high-stakes assessment (in the sense that the results of which can impact upon students’ tertiary entrance scores or access to university) and that is grounded in contemporary constructivist learning theory (Shepard 2000) is the Senior PE programme in Queensland, Australia (Macdonald and Brooker 1997a, 1997b; Kirk and O’Flaherty 2004; Hay 2006). The assessment in Senior PE is school-based and requires that teachers make judgements against criteria and standards of student performances (physical and written) in authentic assessment contexts and tasks. The subject has been celebrated for focusing on more sophisticated and higher order learning in the physical and cognitive domains, employing a criteria and standards referenced approach to assessment, and valuing teachers’ professional judgements and curriculum/assessment construction skills. However, no research has been reported on the appropriation of the stipulated assessment procedures and processes in Queensland Senior PE. This paper investigates the processes that teachers from two school contexts employed to judge the quality of students’ movement performances using a criteria and standards referenced approach to assessment.

*Corresponding author. Email: phay@hms.uq.edu.au
Validity and reliability of criteria and standards referenced assessment

The use of criteria and standards for determining student achievement has gained increasing popularity in western education systems commensurate with a drive towards more performance assessments. These assessments are purported to be more ‘authentic’ than traditional pen and paper exams or psychometrically based measures and give greater opportunity for students to demonstrate complex and higher order cognitive processes (Popham 1997; Shepard 2000; Crehan 2001). This said, validity and reliability, along with comparability and fairness (Linn et al. 1991), are necessary to consider in relation to the employment of any educational measure as they are ‘social values that have meaning and force outside of measurement wherever evaluative judgements and decisions are made’ (Messick 1994, 13). Given the research paradigm employed in this study, it is impossible to indicate statistically the validity and reliability of the criteria and standards approach utilised by the teachers. Nevertheless, the two principles are pertinent to the issues raised (Stokking et al. 2004).

According to Messick (1989, 13) ‘validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment’. Validity is not a principle confined to psychometrically founded tests but is relevant to any process or practice of observing or documenting behaviours or attributes (Messick 1998; Kane 2001). Furthermore, it is not restricted to the internal machinations of the measure but includes the social consequences of the interpretations and uses of the measure (Messick 1989, 1994, 1995; Kane 2001). There have been numerous studies into the validity of performance assessments that usually utilise rubrics (referred to in this paper as criteria and standards matrices). Such studies, for example, have indicated the difficulty associated with collecting relevant evidence to substantiate the validity of these measures (Koretz et al. 1994); questioned the relevance of the content of the tasks and thus the teachers’ decisions (Stokking et al. 2004); described validity variability depending on the type of rubrics employed (Novak et al. 1996); and recommended cautious use of performance-based assessments in high-stakes contexts because of the modest inter-rater agreement and generalisability (McBee and Barnes 1998; Crehan 2001). Hoge and Coladarci (1989), however, concluded that assessments employing teacher judgements were of greater validity than tests because the teachers’ judgements covered a wider range of student outcomes.

Reliability is ‘concerned with the accuracy with which the test measures the skills or attainment’ (Gipps 1994, 67) and relates to the replicability and comparability of pupil performance and the assessment of that performance (Feldt and Brennan 1989). Moss (1994, 7) suggested that ‘as assessment becomes less standardised, distinctions between reliability and validity blur’. Such blurring is endemic to the assessment processes in Senior PE because of its school-based nature, also foregrounding for particular consideration inter-rater reliability of the process. The halo effect, leniency error and severity error are examples of concerns that have been identified regarding the reliability of teachers’ qualitative judgements (Nitko 2001). Low inter-rater reliability scores for performance assessments of writing and research skills have been reported by Koretz et al. (1994) and Stokking et al. (2004), respectively. Similarly, low inter-rater reliability ($r = 0.40$) in the assessment of teacher performance was reported by Bond (1995) who argued that assessors held unarticulated beliefs about what constitutes evidence and that these beliefs were resistant to change even after training and calibration. On the other hand, Harlen (2005) reported, in a review of assessment literature, that the reliability of teachers’ judgements has been demonstrated
to be consistent with the reliability of traditional tests. Furthermore, the presentation of evidence indicating concerns over the validity and reliability of rubric use does not imply the substantiated validity and reliability of traditional assessments.

**Senior PE**

Senior PE is one of a suite of subjects offered in the Queensland senior school curriculum that is characterised by school-based, authentic assessment that contributes to the generation of a tertiary entrance score (Queensland Studies Authority 2004). The assessment regime is based on a portfolio system in which evidence of student learning is collected continuously across the two years of the course. This evidence contained in the portfolios is ‘selectively updated’, acknowledging that evidence collected at an earlier date might no longer be representative of student achievement, and recognises the dynamic nature of learning (Rovegno and Kirk 1995). In this context, schools are responsible for the construction of assessment devices and task-specific criteria and standards reflecting the statutory guidelines and generic criteria and standards matrices of the official syllabus documents. The teachers are required to reference students’ performances (in this case their physical and written performances) against the task-specific criteria and standards.

Sadler (2005) noted that the use of educational terms, particularly those concerned with assessment, are inconsistent in the educational literature. This makes comparison difficult and can create confusion if the usage of terms is not clearly articulated. The assessment approach in Senior PE draws on a representation of criterion-referenced assessment that involves the subjective interpretation of the quality of a student’s work in relation to often quite broad criteria and standards. Sadler (1987a, 193) suggested that this approach to assessment ‘draws on the professional ability of competent teachers to make sound qualitative judgements of the kind they make constantly in teaching’. Specifically, criteria in this form define the aspects of a subject or domain that are to be measured (Bingham 1987). Standards represent the distribution of quality along a specified continuum (from highest to lowest) and articulate that quality for each criterion that the students’ work will be measured or referenced against (Sadler 1987b).

Teachers are required to develop task-specific criteria using this matrix for each assessment task or context. Table 1 represents the Senior PE syllabus standards for physical performance under the ‘Acquire’ criterion.

An ‘A’ standard is the highest standard and carries the greatest value in terms of contributions to exit levels of achievement and tertiary entrance scores. The descriptors of an ‘A’ standard student are listed in Table 2 as they best describe what is valued in this subject.

<table>
<thead>
<tr>
<th>Standard A</th>
<th>Standard B</th>
<th>Standard C</th>
<th>Standard D</th>
<th>Standard E</th>
</tr>
</thead>
<tbody>
<tr>
<td>The student: performs a wide range of controlled physical responses that demonstrate fluency, accuracy and, where applicable, speed</td>
<td>The student: performs a wide range of controlled physical responses that demonstrate accuracy and, where applicable, speed</td>
<td>The student: performs a range of controlled physical responses accurately</td>
<td>The student: performs physical responses</td>
<td>The student: attempts physical responses</td>
</tr>
<tr>
<td>Acquire</td>
<td>Apply</td>
<td>Evaluate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The student:</td>
<td>The student:</td>
<td>The student:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• performs a wide range of controlled physical responses that demonstrate fluency, accuracy and, where applicable, speed</td>
<td>• through analysing and organising a wide range of physical responses of self and others, successfully adapts personal and team performance in team and individual circumstances in complex performance environments</td>
<td>• consistently and successfully initiates and implements physical responses through reflection and decision making in new or unrehearsed contexts within complex performance environments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• demonstrates, through physical performance, a knowledge and understanding of learned rules and sophisticated (complex) rehearsed strategies</td>
<td>• consistently and successfully applies a wide range of team and individual strategies that enhance performance of self and others in complex performance environments</td>
<td>• consistently and independently solves problems by physically demonstrating the solutions in new or unrehearsed contexts within complex performance environments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• recognises, recalls and comprehends a wide range of terminology, principles and concepts specific to the focus area and the physical activity under study</td>
<td>• consistently, successfully and independently compares, categorises and manipulates information</td>
<td>• consistently solves problems and makes decisions to produce a wide range of credible and convincingly justified evaluations of physical activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• demonstrates sustained and accurate use of textual features within the mode of communication</td>
<td>• consistently and successfully uses a wide range of modes and genres to complete the requirements of a wide range of tasks by applying and integrating ideas and information</td>
<td>• consistently provides effective solutions in the presentation of communications that enhance meaning and impact</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These exit criteria are derived from the global aims of the syllabus – Acquiring, Applying and Evaluating – and provide the connection between the learning experiences and assessment expectations. The criteria were originally and essentially based upon Bloom’s (1956) taxonomy of educational objectives and now, also, draw on Marzano’s (2001) taxonomy in relation to the links between cognition and physical performance. ‘Appreciating’, the fourth global aim of the PE syllabus, is not represented in the criteria and standards matrix and therefore information is not gathered in relation to the movement towards, or dispositions pertaining to, the affective domain. The rationale for this absence is not provided in the syllabus. In assessing the physical or movement performances of students in Senior PE, the syllabus encourages the use of progressive judgements across units of work. Specifically it explains that ‘Physical assessment continues through the unit of work. Tasks should be created to help confirm teachers’ decisions. A single-point-in-time task should not be the only mechanism used by teachers to make decisions about students’ physical responses’ (QSA 2004, 48)

The reliability of teacher judgements using the aforementioned approach depends upon three conditions: time to make judgements and reflect upon them; internal moderation of teacher judgement (within a PE department, for example); and external moderation, and supposes that ‘emphasis is placed [by the teachers] on developing criteria and standards that are explicit, well-articulated, well-understood and easily internalised and applied by teachers’ (Pitman et al. 2002, 328, our emphasis). Pitman et al. (2002, 328) further explained that criteria and standards should ‘facilitate clear thinking and communication amongst teachers, enhance the comparability and portability of results, and help students (and their parents) understand assessment decisions’. This is fundamental to both the translation of the generic set of criteria and standards into task-specific matrices and to their use by teachers in classes.

Teachers’ and schools’ compliance with the syllabus requirements for implementation of the curriculum, including assessment is accounted for through state-wide moderation (via district panels) of both the school’s work programmes and assessment practices (including both the design of assessment items and the application of criteria and standards in determining levels of achievement). Moderation is a process for ensuring consistency of judgements about the quality of students’ work and involves teachers comparing students’ work and coming to a shared understanding of the criteria and standards against which the students’ work is measured (Cumming and Maxwell 2004). Moderation is purported to promote comparability and equity in the application of standards. In PE moderation of teachers’ physical performance judgements involves their submission of video examples of typical ‘A’ and typical ‘C’ student responses in two of the four physical activities. It is assumed that the video evidence represents the application of standards for each student in each activity studied.

The study

The data presented in this study are from semi-structured interviews with and participant observations of two teachers at secondary school sites: one, an ‘elite’ coeducational private school (School A), and the other a low socio-economic government school (School B). Interviews were conducted at five specific junctures across two consecutive Senior PE units (each unit is about 10 weeks in length) according to the school’s PE curriculum and assessment plan and the participants received transcripts of their interviews to confirm their accuracy. The teachers were required to identify six students in their Year 11 Senior PE classes (Year 11 is the penultimate year of formal schooling in Queensland) in three broadly
defined ability categories (high, middle and low). The teachers’ identification of these students provided important analytical reference points for both the participant observations and the semi-structured interviews.

Participant observations of one lesson per week were made according to the observation protocols of Sarantakos (1998) and focused particularly on the extent and nature of the interactions between teacher and students, the assessment processes employed by the teachers (as indicated by them), and the interaction between student participants and other students in the class. The observations contributed to the direction of subsequent semi-structured interviews.

Interviews with the two teachers focused on a range of issues including the assessment processes they employed across the two units; the achievement and progress of the students; explanation of and justification for the achievement decisions made in each unit in relation to the student participants; and the nature of the identified students’ abilities. The data has been analysed within a post-structural paradigm to focus on the discursive resources that particular individuals drew on to understand others, and the social and cultural consequences, in terms of power and positioning, of this constitution (Macdonald et al. 2002) for students’ achievement in PE. Following the principles of a constructivist version of grounded theory (Charmaz 2000) the data were systematically coded and analysed in relation to the points of focus described above in an attempt to build a theory faithful to the evidence generated through the interviews and observations (Neumann 2003). To preserve the anonymity of the teachers and schools featured in this paper the participants and schools have been ascribed pseudonyms.

**Results**

The initial data presented here are derived from interviews conducted with Mark, the Head of Department (HOD) of Health and Physical Education at School A, an ‘elite’ co-educational private school. His standing in the broader field of PE in Queensland is indicated by his appointment to a review panel for Senior PE. As a panellist he is responsible for ensuring that the curriculum and assessment tasks constructed by teachers are within the guidelines of the syllabus and that the standards applied by teachers are consistent across schools and with the generic set in the syllabus. The second set of transcript excerpts are from interviews with Anthony, an experienced teacher at School B, a government funded secondary school situated in a low socio-economic area. Anthony had been teaching Senior PE for eight years and believed he had a strong grasp of the intent of the syllabus and had achieved the ‘internalisation’ of the criteria and standards.

As mentioned previously, the use of criteria and standards for making decisions about student achievement depends in part on the extent to which the criteria have been ‘internalised’ by the teachers (Pitman et al. 2002). Furthermore, the syllabus indicates that students’ physical performances are expected to be assessed by teachers progressively across each unit of work.

When you’re doing that it’s a difficult one because you’re doing this progressive assessment and you’re looking at kids all the time and I don’t think you, necessarily… you’ve probably got the matrix in the back of your mind and you have internalised what you think it means, but it’s not what drives your time…. So, sometimes [laughs] I think you’re flying by the seat of your pants, to be truthful. We use our gut feeling, quite often and then I think you go away and you look at each student in turn and then that’s when you try and come back to the matrix and marry up what you’re thinking with what is in the matrix. (Mark)
Mark explained that assessment is progressive but his judgements are not necessarily dependent upon or ‘driven by’ the criteria and standards. His approach to assessment is not dissimilar to the ‘intuitive assessor’ proposed by McCallum et al. (1995) in the sense that he demonstrated a loose reliance on the actual standards descriptors as the basis for the decision made. Such a finding was also indicated by Hawe (2003) who noted a divergence from the official criteria and standards when assessors took into account factors irrelevant to the actual student work, such as behaviour, attitudes and other biographical and contextual factors. Mark’s reference to decisions made on ‘gut feelings’ may indicate an internalisation of sorts; however the absence of practical reference to the criteria and standards during the term and the retrospective reference to the matrix after the judgements were made undermines the probability that the descriptors in the matrix are the sole basis for those decisions. ‘Gut-feeling’ was also important in the way Mark developed perceptions of his students’ abilities, but in this case, he was more explicit about the contributing influences on the perceptions he had of the students.

Well, first of all when you start teaching, when you first get them together, just their demeanour in the classroom for a start, gives you a bit of an indication. Whether they’re switched on; whether they respond to your questions or your teaching; how they respond to each other: whether they’re, you know…. You get a bit of a feel about their level of enthusiasm; are they prepared to contribute to the class? Those sorts of things. It’s fairly easy in the physical side when you put them out on the volleyball court and you ask them to play a game of volleyball, which we do very early in the piece, without too much tuition. We ask them to get out there on court and play a game and we video them and then they assess themselves, too, at that stage. That gives us a pretty good indication, early on, about who’s got past experience and who can do the sorts of things that are required in volleyball, you know. But as far as the academic side of things is concerned and the classroom work it’s a matter of how they interact with you in the class early on. (Mark)

Mark’s perceptions or ‘gut-feelings’ about the abilities of the students seemed to be initially informed by the behaviours and interactions of the students that were indicative of certain valued dispositions (such as being switched on, enthusiastic, interacting with the teacher through questions and answers) to physical activity/education. The significance of these factors was substantiated through Mark’s account of these prior to the students’ engagement in movement contexts. In returning to the processes involved in gauging students’ levels of achievement in PE, it can be seen that Mark similarly referred to affective markers such as motivation and attitude which he was able to observe over time.

Well… I sort of look at the whole… I mean it’s just over time. See sometimes… you look… with our criteria for the physical performance in track and field they… what they do in their training is, makes up part of their training too. So you’re looking at how they put their training together and how they perform the skills in their training and these sorts of things. Because you’ve got so many students you might just see a small part of the work and you build it up over time. And you look at their attitude, their attitude and their commitment and their motivational levels and… I don’t think they’re mentioned in the matrix, but they’re the sort of things that give you a few clues about how they’re going. (Mark)

The valued affective characteristics such as commitment and motivation that are displayed across the unit by students influenced Mark’s formation of his perception of their standard against the criteria even though they are not represented in the descriptors. This is not to say that the PE activities and content knowledge of the syllabus and represented in the matrices were not considered, but rather that the affective characteristics of a student contributed to, or ‘filled out’, Mark’s perception of their ability, performance and standard.
The valuing and incorporation of such factors (informed by Mark’s expectations, beliefs and values) represent content irrelevance, and by implication construct irrelevance (Messick 1995), as they are not specified within the boundaries of the construct domain that was being assessed. The poignancy of the students’ characteristics, the behaviour of the student, and the extent to which they have engaged in meaningful or valued ways in the pedagogic context, seemed to be important when Mark determined the final grade for the unit, further raising questions over the construct validity of the decisions made and the criteria he employed to make those decisions.

No, it’s basically what you can remember. I thought about doing checklists and things like that but there’s so much going on and there’s so many students doing so many different things that you never teach. Once they start the training programme [for athletics] almost all the teaching at that point is individual. I mean sometimes you would take the video camera there and you’ll, you’ll set aside certain lessons when um you try and video something that most of them are doing so you’ve got a little bit of fall-back there. But, generally speaking, it’s what you’ve got locked in your brain. (Mark)

This idea of having a perception of a student ‘locked in your brain’ is important and implies that the students have had the opportunity to display the valued performance attributes (whether they are those articulated in the criteria and standards or those unique to the teacher’s values). This notion was developed further by Mark.

I think that most of us make our judgements first and then we look at the criteria sheet and see if it backs us up, although I’m going to try and go the other way this time. Although, I guess you have in the back of your mind, generally, where the student is, you know. You do because you’ve seen them do what they’re doing for so long and you think ‘yeah, well, they’re about here’. So, you just can’t avoid that. (Mark)

Although some video evidence was collected during the units (principally for the purpose of moderation), memory, it seemed, provided the most useful and readily available evidence upon which Mark’s decisions for each student in the cohort were made. Such an approach to the collection of evidence fails to satisfy the structural aspect of validity which Miller and Linn (2000, 370) indicated required ‘the implementation of rigorous and systematic scoring procedures’, and further confirms the intuitive nature of Mark’s approach. Moreover, the contributing factors to the teacher’s development of memories of students are significant to the achievement potential of those students. Mark’s suggestion that he had made his judgements before explicit reference to the criteria and standards indicates that he already had an idea in his mind about the constitution of quality performance. Furthermore, for students to be ‘remembered’ they needed to perform in noteworthy ways, i.e. ways valued by the teacher. Thus the characteristics of noteworthy performances may not necessarily have been confined to aspects of a particular physical activity of focus, but also the mode of students’ engagement in the physical activity.

The influence of students’ physical, social and cultural resources on the teacher’s perceptions of abilities and approach to assessment seemed evident in the second context of the study. While it is difficult to determine the extent to which one’s perception drives the decisions that are made, there is little doubt from Anthony’s comments that he has expectations of the achievement level of the students in relation to his perceptions of their ability. While ability may have construct relevance (Kane 2001), Anthony’s perceptions had been based on construct-irrelevant characteristics such as aggression and competitiveness. In response to the first author’s question of how the class went in terms of their achievement in term two Anthony explained:
Anthony: I think they went, you know, reasonably well but I'd sort of like to look at my results to reply to that. I know the kids I expected to do well, did, and the kids who didn't or weren't expected to do well basically from memory I don't think they were like standouts.

Peter: How did that expectation get developed?

Anthony: I guess… as I said previously because I've known the kids and I've seen their work in the previous assessment pieces. So I know what they're capable of. So, I guess through observation.

Anthony’s suggestion that his expectations of students’ achievement had been shaped by their previous performance is not particularly surprising. In fact, Trouilloud et al. (2002) suggest that teachers’ expectations may be confirmed because the teachers are competent in predicting, without influencing, student achievement. However in the light of Sadler’s argument for criteria and standards referenced assessment that students’ grades should not be contaminated by the quality of work of other students and should be determined ‘without regard to each student’s previous level of performance’ (2005, 178), it does raise questions about the potential contamination of the judgements made in this unit.Anthony’s reference point for his decision-making was similar to Mark’s, but there was a more explicit reference to the observation of a student’s ability over time as an informant to the assessment process. The opportunities for students to display the valued characteristics (both those represented in the syllabus and his situated expectations) were tempered by the contexts that Anthony provided for the students to perform in.

The way I grade them often is I actually play them myself and I do my best and I expect them to meet the challenge if they can. (Anthony)

This perhaps seems a useful way of getting an indication of the standard of each student’s physical performance. The teacher is able to see each student in a controlled context and assess both skill proficiency as well as the decisions made during the game. Yet, in practice, it was observed that Anthony tended to play against particular students who wanted to compete with him. Consequently, these students spent more time on court with Anthony than other students who were physically and socially less confident. Besides the obvious equity concerns related to opportunity for interaction with the teacher, Anthony’s assessment practice could conceivably have facilitated a problematic reconstitution of conceptions of quality. In this instance his perceptions of certain students’ performances (the performances of those who were proactive in their engagement of Anthony in a game) rather than the criteria and standards could have become the reference point for grading the other students in the class (Wyatt-Smith 1999). Although this is a somewhat speculative point, Anthony did, at various times, use the attributes of certain students to substantiate his justification for the achievement of others.

They [Chantel and Jodie] didn’t really achieve, um, marks that were anywhere near high level. I’d say they’d be even close to um… I think I um ended up giving them Cs or low Cs, high Ds, and again I think that comes down to motivation. They don’t appear to be… they definitely don’t have the, the want to achieve as, the desire to achieve as Emma and um Steven do. (Anthony)

In challenging Anthony to a game, students displayed the attitudes and dispositions that were valued by him. This was reflected in his comments about the ability and achievement potential of the students.

Anthony: Looking at his prac marks, he’s got straight ‘A’s for his Acquisition, straight ‘A’s for his Application, ‘B’s for his Evaluation… sorry… ‘B’, ‘A’, ‘B’. First of all,
Steven has fantastic skills, not looking at the matrix at all, he’s got very good skills. He has very good hand-eye coordination, obviously, great speed. He seems to have a drive to win as well; very aggressive in everything he seems to do. Because of that, once again… he practices hard and he … I think he’s ironed out a lot of his mistakes. He’s fairly consistent.

Peter: Did that help you make those judgements when you were applying those standards?
Anthony: Um, yes, definitely.

References to the criteria include the speed and consistency of the students’ movements in the performance context. However, the justification for the grades also included reference to Steven’s ‘drive to win’ and aggression in play. ‘Drive’ reflected Anthony’s own statement about doing his best when playing against the students, and his expectation of students to meet that challenge. It seems that Steven, through particular actions demonstrated not only the physical and strategic practices that were valued in the syllabus, but also the attitudes and behaviours that were valued by Anthony. Steven’s opportunity to display and receive recognition through assessment was possible because of his proximity to the teacher in game contexts. Once again such influences on the teacher’s judgement of a student’s performances indicate construct-irrelevance and raise questions about the validity of the decisions made.

In relation to Emma, identified by Anthony as a high ability student, and Chantel, identified as low ability, his justification for Emma’s strong progress and Chantel’s comparatively mediocre achievement was similarly described in terms relating to affective characteristics. His perceptions of Chantel’s motivations, in particular, were based less on observation given his limited interaction with her.

Emma [a student identified as ‘high ability’] is certainly progressing well. I worked last week with Emma in some strategies, problem solving activities and you know she just has the right sort of aptitude to handle it… doing well. As far as Jodie and Chantel are concerned I haven’t worked with either of those two girls at this stage. Um... I think to some level they’re still working on their skills. You know, they haven’t started to look at game strategies and that sort of stuff, yet. (Anthony)

Emma, a student Anthony describes in similar terms to Steven, benefited from sustained interaction with Anthony. Furthermore, Anthony, in this situation, was able to assist Emma develop her higher order processing skills valued in the ‘Evaluate’ criterion through his personal interactions with her and specific feedback on strategic awareness in the game. Conversely, the lower ability students had comparatively (and at this point absolutely) less opportunity for interaction with Anthony. Later in the interview Anthony was asked, how much difference a change in motivation would make to Jodie and Chantel’s potential to achieve.

If they were super motivated they would achieve, obviously… I think they would achieve higher marks because being a motivated person they’d be in there practicing or asking questions, making sure they get on the court and getting as much court time as possible. So... as a lower achiever, you expect the opposite of that to occur. You know they’re just… as I said earlier, they just don’t get out there and they don’t ask questions or they don’t get out there, they don’t do any extra work. (Anthony)

Anthony’s perception of Jodie’s and Chantel’s motivation was based on their supposed unwillingness to ‘make sure that they get on court’, ask questions and do ‘extra work’. It was observed that while Jodie’s engagement in the performance context was sporadic, Chantel was involved in every lesson but spent many lessons playing on a make-shift court
in the wings of the hall while the more proactive (or motivated, as described by Anthony) students played on the courts proper. The courts were spaces where feedback and assistance could be given and in which recognition of valued attributes could be realised. Chantel was not given equal access to these spaces but was nonetheless graded as if this was the case. Furthermore, because of Anthony’s situation in that space, he was blind to his own assessment prejudices.

Discussion

Linn and Gronlund (1995) suggest that the value of criterion-referenced assessment of complex performance domains is that both the product of the students’ experience and the processes resulting in the product are accounted for without influence from the achievement standards of others. Such an approach assumes that the judgements made relate entirely to the students’ performances of interest and reliably capture those processes in the assessment contexts. In all units within this study the teachers drew on progressive observations across the term to assess the students’ physical performances. Their referencing processes depended on the ‘internalisation of the criteria’ so that meaningful observations and judgements about the students’ ‘Acquisition’, ‘Application’ and ‘Evaluation’ of unit-specific subject-matter and processes could be made. The teachers understood that this internalisation was important for making ongoing or progressive judgements of students’ physical performance across the term and suggested that such an internalisation had occurred for them.

The assessment practices were not inconsistent with the approach proposed in the syllabus. Yet when the teachers were asked to reflect on the reasons for their assessment decisions, the valued affective characteristics (attitudes and behaviours) and sporting histories of the students were significant in their justification. In effect, the internalised criteria and standards served as alternative criteria and standards that were constituted by elements of the official documents as well as the teachers’ values, beliefs and expectations. As such they were qualitatively and substantially different from the syllabus or task-specific criteria and standards supposedly used. The teachers’ ‘gut-feelings’ about the standard of the student performances were moderated by the students’ possession of contextually valued affective characteristics. The poignancy of these feelings and the clarity of a teacher’s memory of students depended on students’ access to contexts in which the valued resources (actual performance and the supporting attributes) could be displayed. Furthermore, the retrospective referencing of students against the matrix on the basis of memory and already formed judgements distanced the students from the decision-making process itself. The incorporation of these extraneous or construct irrelevant factors, as well as concerns over the structural validity of the collection and collation of evidence (through memory) by the teachers indicate that the construct validity of the decisions and the teachers’ assessment practice was questionable.

Concerns with inter-rater reliability potential are therefore also highlighted as reliability should depend upon a shared and consistent reference to the explicit standards descriptors of the syllabus. This is inherently difficult in relation to the notions of progressive assessment and internalised criteria as reliability will depend on the extent and accuracy of this internalisation and the consistency of teachers’ expectations, values and beliefs in relation to PE and students. Admittedly, this paper has not addressed the process of moderation aimed at ensuring that the standards are applied appropriately and consistently by the teachers within and across schools. Irrespective of the power of this process to promote the reliability of the approach (Maxwell 2002), it cannot address the initial decision-making
processes of the teachers or the implications of this decision-making-process on students’ access to ‘spaces’ within the pedagogic context for recognition of learning. Validity, and implicitly reliability, are matters of degree, not all-or-nothing properties of educational measures (Messick 1989; Kane 2001). We are not arguing that the teachers’ approach was invalid but that a degree of invalidity and unreliability was evident. Given the high-stakes nature of the assessment and the implications for students’ possible futures, beliefs about themselves, and future engagement in a movement culture, we suggest that the findings are cause for concern.

In numerous ways, the assessment practices demonstrated by the teachers in this study reflected the intuitive assessment model proposed by McCallum et al. (1995). For example, the teachers relied on memory as evidence upon which assessment decisions were made; there was a sense of teachers possessing their own implicit standards; perceptions of the students’ attitudes and behaviours were incorporated in the teachers’ assessment of students’ performances; and the teachers indicated that their knowledge of the student strengthened their ability to assess them. Yet there were some important differences. For example, the teachers in this study were comfortable with informal observations as the basis for the awarding of movement performance grades at the end of a unit. Furthermore, they did not believe that they were minimal adopters of the assessment procedure but rather were assessing students in a manner consistent with syllabus requirements. Although Mark did indicate a belief that more formal collection of evidence across the unit was desirable, neither he nor Anthony indicated that their practices were contrary to official expectations. In the light of this, it would seem that the notion of ‘internalisation of the criteria’ is confusing and, in connection with the syllabus requirement for progressive assessment, promotes an intuitive approach to assessment that is somewhat problematic in terms of validity and reliability.

While we have presented evidence that raises questions of the validity and reliability of teachers’ assessment of physical performances, we believe that alternative assessments that employ teacher judgements are educationally valuable, and that high-stakes assessment in PE is worthwhile. In order to ensure its viability attention needs to be given to the nature and clarity of the official assessment guidelines that teachers are provided with (specifically what and how progressive judgements are to be made and accounted for), and to processes beyond external moderation to ensure the teachers’ valid and reliable appropriation of the approach. Transparency in the decision-making process of teachers needs to be further promoted (beyond external moderation) to maximise the fairness of the approach. Transparency may be promoted through the statutory requirement of teachers to give greater justification to the students for the referencing decisions that are made. This includes students in the decision-making process and promotes the monitoring of their own learning and achievement.

Greater clarification and detail is needed as to the nature and expectation of progressive judgements and the internalisation of the criteria. We believe that such principles should not be promoted at the expense of the requirement for the systematic collection of meaningful and defensible evidence of student learning. Such evidence could be collected through video footage (which is already collected on a much smaller scale and for relatively fewer students for the purpose of moderation) and recorded observations and checklists. This will, in our opinion, reduce the reliance on memory as a source of evidence for decision-making, which seemed to be tainted by construct-irrelevant affective information and mitigated by students’ access to contexts for recognition. Although we acknowledge that such a requirement is burdensome in terms of teachers’ time in the classroom and beyond it, we believe that it could promote greater access for students to both the constituting expectations of the
‘internalised criteria’ and consolidate the official elements of the criteria and standards as the reference point for referencing decisions.

Rigorous training procedures might also be instituted to promote the optimisation of teachers’ assessment comprehension and competence. In achieving the reliability of teachers’ assessment of student writing Brown et al. (2004) engaged teachers in a comprehensive training process that included, among other elements, instruction on content; an overview of the scoring rubrics; detailed and specific training clarifying the task to be assessed; use of sample scripts as benchmarks; cross-checking of marking by expert panels. We similarly believe that teachers benefit from specific, intensive and somewhat regular training and engagement in marking panels. This should involve instruction on the collection of evidence for decisions, cross-checking of teachers’ referencing against control material (in the form of videos and written work), and collaborative assessment with trained markers to help consolidate the reliability of decisions. All teachers, rather than those involved in moderation, need to be skilled in the use of criteria and standards and given access to the professional dialogue that such training could promote. We believe that the training should also include the promotion of teachers’ reflexivity to address the source and nature of the construct-irrelevance highlighted in this study. In this case we refer to the self-analysis of one’s beliefs, values, history, etc. in the recognition that these aspects unavoidably influence one’s perceptions of others and interactions with them. This approach may aid teachers in their understanding of the influence they have over student achievement and opportunity to achieve and help them develop pedagogic and assessment practices that address the problematic consequences of such aspects.

Conclusion

Student achievement grades or marks serve, in one sense, as symbols of scholastic quality. These symbols provide a means of certification and selection in educational institutions because there is a presumption that a shared understanding of the contributing elements of the symbol exists amongst education stakeholders. This is particularly claimed in relation to criteria and standards-referenced assessment because the descriptors should articulate the basis upon which judgements are made. This study has suggested that in Senior PE, where assessment is progressive and the subjective decisions are difficult to provide evidence for, the constituting elements of the symbols awarded to students are somewhat elusive. The suggestion that reliable and meaningful judgements can be made when the criteria and standards have been sufficiently internalised by the teachers has intuitive appeal, however the data gathered for this study suggests that in fact such an internalisation creates a new set of criteria and standards that bear some semblance to the official set outlined in the syllabus, but which are embellished by teachers’ values, beliefs and expectations of students and the subject (in this case, PE). Furthermore, this ‘new’ set of criteria and standards are not as accessible to the students, so the students are more disconnected from the decision-making processes.

The concerns raised in this paper regarding the assessment approach employed by two experienced teachers in the Queensland Senior secondary education system are particularly pertinent because of the summative and high-stakes nature of the assessment (Stokking et al. 2004) and the possible consequences of the students’ results for their educational outcomes (Messick 1994, 1995; Lubienski 2003). Our interpretation of the data indicates that educational possibilities for students are potentially undermined or unfairly promoted by institutional practices that should be about ‘both quality control and the due recognition of achievement’ (Sadler 1987a, 196). Data indicate that students’ achievement may be
affected by alternative structures that potentially enhance the achievement possibilities of some and marginalise others. Further research on a broader scale is needed to determine the extent to which the processes identified in this study are occurring in other curriculum contexts. Research is also needed to assess whether the solutions to the problems require changes to the assessment approach or whether enhanced professional training, development and support may better ensure more valid and reliable referencing procedures in the performance dimension of PE and, potentially, of other subjects too.

Notes on contributors

Peter Hay is an associate lecturer in the School of Human Movement Studies, The University of Queensland. Drawing largely on qualitative methodologies, Peter’s research interest focuses on ability construction and assessment practices and outcomes in physical education and other learning domains in the broad field of education.

Doune Macdonald is a professor and head of the School of Human Movement Studies, The University of Queensland. Her research interests have focused on the challenges of curriculum reform and its impact upon teachers and teaching in health and physical education through the lens of professional socialisation, discourse analysis and identity construction using predominantly qualitative methods.

References


