Journal of Maltese Education Research

Vol:1 No.2 2003 15-34
www.educ.um.edu.mt/jmer

Validation of Written and Video Based Assessment Instruments in Physical Education

Gemma van Vuuren-Cassar

gemma.vanvuuren-cassar@um.edu.mt

Dr Gemma van Vuuren-Cassar has been a lecturer at the University of Malta since 1995. She is attached to the Faculty of Education and the Institute of Physical Education and Sport and she is primarily involved in teacher training. She read for a Diploma in Applied Social Studies (1989) and a Bachelor in Education (1991) at the University of Malta. She has taught in Primary Schools and Physical Education in Secondary Schools. Following that she read for a Masters in Education (1995) and a PhD. (2000) at the Victoria University of Manchester, England. Her doctoral thesis focused on the teaching and assessment of physical activities at GCSE and AS levels. Her areas of research and publications include curricular issues, pedagogy, assessments and examinations in physical education.

Iasonas Lamprianou

iasonas.lamprianou@man.ac.uk

Dr Iasonas Lamprianou joined the Centre for Formative Assessment Studies CFAS, University of Manchester (1999) as an analyst after experience as a primary school teacher and researcher at the division of Research and Evaluation of the Pedagogical Institute in Cyprus. He has participated in research projects involving educational assessment, measurement, test equating and Item Banking. His research interests include psychometrics and especially test-equating, Item Response Models and person-fit statistics. Currently, he is mainly involved as an analyst in the School Sampling Project (SSP) and he contributes to the Standards and Evaluation Reports for the National Curriculum (NC) tests. He is also generating diagnostic feedback for schools based on scrutinized NC test results.

Abstract:

A written paper (WP) and a video-based-paper (VP) were used to measure the learning outcomes of three experimental teaching units (ETUs) of athletics. The subjects (n=49) were 16 year olds. The ETUs represented three teaching environments; practice, practice and handouts, and class-based sessions.

Pre and post-tests were administered for the content of rules, planning tactics and techniques. Rasch model analysis showed that both tests were unidimensional and reliable (R ranged from 0.86 to 0.92).

The efficiency of the three experimental teaching units was evaluated by means of paired sample t-tests. The subjects performed better on the post-tests (WP: effect size=1.2 and VP: effect size=0.93). The pre-tests and post-tests abilities of the subjects were highly correlated (WP: r=0.380 and VP: r=0.322). The subjects of the class-based sessions achieved significantly better scores on all content areas when video based assessments were used. The findings of this study are applicable to athletics.

University of Malta, Msida, Malta MSD06
ISSN **1726-9725**

Introduction

This study is on the teaching and assessment validation of the practical performance coursework of examined physical education in secondary schools. Practical performance coursework of physical activities include content such as rules, planning tactics, techniques, methods of training and health related exercise. The assessment instruments of cognitive aspects of this knowledge include written, oral and video based activities. British examiners have been recommending class-based sessions in addition to practice-based sessions (MEG, 1993) for performance coursework. Meanwhile some teachers used homework for the analysis of performance of physical activities (Forsyth, 1994). Nevertheless, external examiners (AEB, 1997) reported that the observation and analysis component of physical activities still needs teaching refinement and rethinking. Although the suggestions of examiners seem to have been taken on by teachers, there still resulted unsatisfactory learning outcomes. There are various modes of teaching and assessment of performance coursework in physical education. Currently, it is still not clear which are the teaching modes that yield the best achievement, especially when different modes of assessment are in use.

This 'study' attempts to investigate this lack of clarity. Dominant modes of assessment of performance coursework content such as techniques, planning tactics and rules will be questioned in relation to different teaching environments namely, practice, practice and handouts, and class-based sessions. This study focused in particular on the teaching and assessment of selected content of athletics, namely sprints, relays and the shot put. More specifically, the study aimed at:

(i) establishing a process of constructing and evaluating three experimental teaching units of instruction and two reliable assessment instruments;

(ii) investigating if the teaching of content (i.e. techniques, planning tactics and rules) in a practical performance teaching situation adequately prepares students for video based and written assessments;

(iii) determining if written and video-based assessments are sensitive to the teaching environment (i.e. performance-based vs. class-based)

Background

The following paragraphs provide a brief overview of the shift from recreational to an educational philosophical base in physical education. This included the use of teaching resources and video assisted instruction. Teaching environments were structured to address the "performance" and "cognitive" aspects of physical activities to be assessed. The impact of internal and external modes of assessments on these developments will be discussed.

Assessments in PE

The introduction of assessments in Physical Education (PE) in secondary schools in the United Kingdom led to a shift from a recreational to a more educational ideological base (Carroll, 1994). This resulted in new strategies, such as teaching for understanding (Thorpe and Bunker, 1986; Fleming, 1994). This emphasised the integration of planning tactics, techniques and rules. The development of the content and modes of assessment of PE at 16+ and 18+ in England followed (Alderson, 1988;

Aylett, 1990; Carroll, 1990a, 1990b, 1994; Francis, 1990). Meanwhile, the direction of the PE curriculum called for instructional designs for teaching physical activities (Vickers, 1990; Melograno, 1996), with a strong emphasis on content (Armstong, 1996) and cognitive abilities such as evaluation, application and analysis of performance (Carroll, 1994; Piotrowski and Capel, 2000). Nevertheless, the external examiners reported (AEB, 1997) that the observation and analysis component of physical activities still needs teaching refinement and rethinking.

Statements of this kind questioned the validity and reliability of unseen written papers, practical performance and oral tests for assessing the cognitive domains of knowledge recall, application, analysis and evaluation of physical activities. The unseen written papers (pre-1998) addressed an arbitrary selection of cognitive and content components, which made the tests unbalanced (Cassar, 1995). Nevertheless, a suggestion to "refine and rethink" teaching strategies raised questions of how content and cognitive domains were presented to the learners. Had teachers considered, if not reverted to classroom-based teaching to beef up the content / cognitive component?

Teaching resources

The availability of teacher resource packs, student textbooks, videos and interactive CD-rom recommended for assessed PE in secondary schools has witnessed a significant expansion over the past two decades. Some recent examples include Beashel *et al.,* (1997); Honeybourne *et al.,* (2000), (2002); and Scott, (1999), (2001). It is interesting to note that most books included a substantial amount of tasks for students. These books and resources cover relevant content such as sport and social sciences; however, they provide limited coverage of tasks referring to the practical performance coursework content.

Roscoe (1998) reviewed various sports activity specific books series such as "Know the Game Series", "The Skill of the Game Series", "Steps to Success Series", etc. These appear to be of a generic nature and include a selection of rules, skills, tactics or umpiring. What is of immediate concern is the absence of practical coursework and sport specific textbooks and teaching resources that address cognitive components such as the analysis, evaluation and improvement of performance. The struggle to innovate the infusion of content and cognitive domains in PE was already reported by Stirling and Scott (1989), and it seems that very little has been reported after that.

Video assisted instruction

The advances in sport sciences have promoted the analysis of motor skills through the use of computers and videos analysis systems such as Dartfish-DartTrainer, Noduls-The Observer, and Kandle- Visual Analysis Systems. Meanwhile, the majority of teachers use a qualitative model to analyse movement performances (Eckrich *et al.*, 1994). Hoffman (1974) described a hierarchical model used to analyse tasks. This includes the skills of observation, evaluation and diagnosis. Pinheiro and Simon (1992) supported the need for skill analysis training in live and simulated environments. However, the majority of training programmes aimed at developing the skill analysis process used the video environment (Wilkinson, 1991; Morrison and Reeve, 1988).

The advantage of videotapes is that they are repeatable, accessible and can include a wide variety of movement responses. Ignico (1994) found that videotape directed

instruction contributed to knowledge, performance and assessment of fundamental motor skills of undergraduates significantly better than teacher directed instruction. Eckrich *et al.* (1994) concluded that video training improves video observational skills and suggested separate training for live performances. On the other hand, Chung (1992) did not find significant differences between a group that received two hours of video interactive instruction as opposed to the same time of traditional lecture presentations for tennis psychomotor skill analysis for undergraduates, in addition to a practice-based course. He also reported that there were no differences between the groups for a written knowledge test, a video analysis test and a performance test. It is observed that while all these studies were concerned with undergraduates, the skills of observation, analysis and evaluation are fundamental for the assessment of the 16+ and 18+ PE syllabuses, and thus have implications for teacher training and teaching.

The expansion of technology in physical education (Silverman, 1997) has been characterised with four categories of potential problems;

1. "have vs. have nots;
2. physical activity (performance) vs. teaching via technology;
3. pedagogic considerations: Is this the best way to teach?; and
4. implications for teacher educators."

The use of video and computer as a replacement of a teacher in a practical session needs to be questioned in relation to the purpose of the technology use and whether it will promote skilled, healthy, enjoyable physical activity. Silverman (1997) has negatively valued using technology for demonstration instead of the teacher, however this generalised statement does not do justice to the different sports activities, some of which require top fitness and skill from teachers to demonstrate (e.g. triple jump in athletics). For the purpose of evaluation and analysis, video recordings of performance are a very important source of data for both athletes and the subjects who are learning and being assessed on these cognitive domains.

Teaching and Assessment of Performance coursework

The structuring of the classroom environment so that students respond to motor learning with greater appropriate practice (Silverman *et al*., 1998) and the significant relationship between appropriate practice trails and achievement have been substantiated in the literature (Buck *et al.*, 1991). However, teachers need to consider student perceptions and the learning environment, because both affect the achievement outcome (Lee, 1997). Sweeting and Rink (1999) suggest that an appropriate environmental design helps students elicit an initial movement pattern, and is also a factor in initial achievement outcome. It is observed that when talking about the teaching environment there is still a heavy notion attached to the practical performance component. Nevertheless, the importance for students to be cognitively engaged in their own learning (Harrison *et al.*, 1999) is gaining acceptance.

The content knowledge of physical activities is the focus of "performance proficiency" and performance-related knowledge, a domain of the disciplinary mastery value orientation of the curriculum (Jewett *et al.*, 1995) that is highly valued in PE syllabuses and textbooks. Nevertheless, Siedentop *et al*. (1994) noted that content knowledge as an objective for teaching PE rated low in priority. If one considers skill improvement, strategic play and acquisition of knowledge, then many of the class ecologies achieved modest learning gains at best. Instruction did not have

the intensity (by which they meant students seriously and intensively engaged in learning activities) other than a literal and symbolic sense of "gym was no sweat!"

Content domains of physical activities were often assessed using multiple choice type questions, often called knowledge tests. These covered content items such as techniques, rules and strategies (Harrison, 1999: volleyball); knowledge of rules, player positions and terminology (French and Thomas, 1987: basketball), and a written test on rules, strategies and mechanics (Rink *et al.,* 1986: volleyball).

Meanwhile, the use of the practical performance tests have been criticised for the lack of opportunities that they provide for the candidates to demonstrate their knowledge and application of the content (techniques, tactics and rules) (Carroll, 1994) and examiners have also commented that

> "the drills seen were not seen as sufficiently challenging and as a consequence candidates were rarely stretched or placed in situations which called for the tactical knowledge or techniques required for a level 7" (ULEAC, 1997, pg. 7).

This implies that the opportunity for candidates to demonstrate their level of performance during the exam is often hindered. The examiners also argued that

> "differentiation in drills is very important thus allowing the more able to spend time on the required advanced skills whilst the least able would not be expected to demonstrate those skills which they have not mastered. The increased use of time spend on the application of the skills… would improve the overall level of skill which is being demonstrated without effecting the length of the session". (ULEAC, 1997, pg. 7).

This rational highlight the feasibility problems of practice-based assessments of an "application" component when candidates had different abilities. The fact that the performance levels of the better or least able candidates had a direct effect on the grading level needs to be considered in the light of the reliability of this method of assessment.

The teacher developed and administered oral assessments of the cognitive components (observation and analysis) have been criticised by the examiners (AEB, 1996 pg. 17; AEB, 1997, pg. 17) for being "restrictive" as teachers did not ask open ended questions and instead asked specific, technical or closed questions tending to lead candidates. Examiners advised that specific questions should be used only if it is felt that the candidate is not showing enough depth in his/her answers (AEB, 1996 pg. 17; AEB, 1997, pg. 17). These kinds of criticisms lead one to question the reliability and the validity problems that are associated with the development and administration of performance coursework and teacher constructed assessments. Thus, the modes of assessing the content and cognitive domains of physical activities through unseen written papers, in the performance setting and through oral assessments were questionable in terms of validity and reliability. Therefore alternatives and modifications of these modes of assessment were necessary to attempt to construct sound assessment instruments, which enable evaluating a balanced variety of content and skills. The process of constructing and validating modes of assessment needs to be investigated in the light of establishing procedures that result in modes of assessment that are valid, reliable, clear and include a balanced variety of the content

of the physical activities *per se.* This study was conducted within the theoretical contexts of emergent theories on instructional design for teaching PE (Vickers, 1990), teaching for understanding (Thorpe and Bunker, 1986) and assessing PE (Carroll, 1994).

Methodology

This study was conducted in three phases. The first involved the development, validation, revision and revalidation of the documentation to be used for the experimental teaching units (ETUs). The documentation included the teaching and assessment objectives, course content, and the two assessment instruments. The second phase involved the development of the lesson notes, teaching activities and resources for each of the three ETUs. These were compiled in a way that the same content was taught in each ETU, addressing all teaching objectives, content, and assessment tasks. Finally there was the delivery of the ETUs and the marking of the pre and post-test scripts by trained markers. The ETUs were filmed for the purpose of systematically analysing teacher and pupil behaviours; however this will not be discuses in this paper.

There were three experimental teaching conditions/environments represented by three ETUs: practice (PE-Athletics class n=12); practice and handouts (PE-Athletics coursework class n=25); and class-based sessions (Sport Studies-Athletics class n=12). The class-based sessions involved discussions, expositions, video clips, handouts and written class work. Each ETU had a duration of 6-lessons of one hour each. Three classes from three different schools were assigned to each of the experimental treatments, while one teacher delivered the three ETUs. The subjects (n=49) were 16 year olds in their first term of post-secondary education in Malta.

This study maintains that performance assessments aimed at assessing the abilities to demonstrate and apply skills, techniques, tactics and rules to the game/competitive situation should be kept separate from assessments aimed at assessing cognitive skills of knowledge, evaluation and analysis.

The construction of teaching objectives, content and assessment instruments of the ETUs were based on PE and sport studies syllabuses namely, General Certificate of Secondary Education (GCSE, 16+ accreditation), Advance Subsidiary (AS) level (17+ accreditation) and Advanced ('A') level (18+ accreditation); athletic textbooks and videos for rules, planning tactics and techniques of sprints, relays and the shot put (Carr, 1999; Roscoe, 1996; BAF, 1994; IAAF, 1990; Pocock, 1995, (*pg. 81-144*), Walker, 1987); cognitive taxonomies of education (Bloom, 1956, cited in Linn and Grondlund, 1995), and past papers of GCSE and 'A' levels. The *Table of Specifications* (Ebel and Frisbie, 1991) procedure was used to bring the course content and the teaching and assessment objectives together. The development of the table of specifications was used to indicate the relative emphasis to be given to each objective and each content area, and to identify the percentage marks which were to be allocated to the content and cognitive domains. The content domain of techniques was awarded 40% of the marks while planning tactics and rules were each allocated 30%. The six cognitive skills (Bloom 1956) were grouped into three categories and each

was awarded a percentage mark loading, namely, knowledge recall and comprehension 30%, application and analysis 40% and synthesis and evaluation 30%.

Table 1: Table of specifications

CONTENT DOMAIN	***COGNITIVE DOMAIN***			
	Knowledge and Comprehension	**Application and Analysis**	**Synthesis and Evaluation**	**Total mark**
Techniques	12	16	12	40
Planning Tactics	9	12	9	30
Rules	9	12	9	30
Total mark	30	40	30	100

Following this exercises a number of questions and their scores were assigned to each cell. The test items used in the assessment instruments were short answer type questions. Diagrams were used in the WP while video clips were used in the VP.

The validation of the ETUs' objectives, content and test items were completed by expert-teachers as recommended by Vickers (1990). The criteria for the selection of expert-teachers required involvement in teaching and coaching (school team or outside school) for a period of a minimum of five years and having attended at least one officiating or coaching/teaching course in the last five years. Five expert teachers were engaged in the validating exercise. This involved rating each item on a set of predefined criteria (Linn and Grondlund, 1995; Ebel and Frisbie, 1991) on a five point Likert scale. Every instructional/assessment objective was rated on two criteria: appropriateness and technical soundness while the whole list of objectives was rated for completeness and feasibility. The validation of the assessment instruments involved the rating of each test item (question and answer) against four criteria: relevance, clarity, correctness and technical soundness. The video clips of the video-based unseen written paper were evaluated for appropriateness and technical soundness by two expert teachers, such that no unintentional information was provided to students. Each test item (question and answer) was also subjected to a validation of the content and cognitive classifications. The validation of the content (technique, planning tactics and rules) and tri-partite cognitive domains (knowledge and comprehension, application and analysis, synthesis and evaluation), required of experts to read each test item (the question and answer) from the provided marking scheme and to tick one of the three content domains and one of the three cognitive domains. The experts were presented with a copy of two tables taken from Linn and Grondlund, (1995, pg. 534, 535) describing the major categories of Bloom's cognitive domain, the relevant objectives and the illustrative verbs stating specific learning outcomes. The validation of the cognitive category domain was anticipated to be somewhat problematic, for various reasons. Firstly, the illustrative verbs provided by Linn and Grondlund (1995) were not exclusive to one cognitive classification (e.g.

describe- used for knowledge recall (level 1) and evaluation (level 6)) and the authentic test items taken from past-papers did not include the same verbs.

Ebel and Frisbie (1991) stated that although a cognitive taxonomy is useful for classifying objectives in terms of level of behaviour required, it is much less useful for classifying test items as it is very difficult to pinpoint the mental processes involved in answering a particular test question, even when we know what took place during instruction. To address this problem Ebel (1965) created a categorisation system to be used with test items instead of objectives, a system which depended on observable operations instead of mental processes. Ebel's Relevance Guide included seven categories: terminology, factual information (corresponded to Bloom's (1956) knowledge), explanation (corresponded to Bloom's (1956) comprehension), calculation, prediction (corresponded to Bloom's (1956) application), recommended action and evaluation (also in Bloom's (1956) evaluation). Nevertheless this classification was not directly useful for the study since the instructional objectives stated explicit cognitive content outcomes.

The first mode of assessment was an unseen written paper with diagrams (WP) while the second assessment involved the attachment of video clips to the questions of the WP. The WP took 30 minutes and the VP 45 minutes to complete. The WP was always administered before the VP. This was necessitated because although the video clips were validated for appropriateness and technical soundness, they were likely to provide clues to the answers, thus contaminating the experimental assessment condition of the WP. For the VP, students were shown the questions and the video clips twice at the beginning, and once at the end of the assessment session. Students were provided with a copy of the questions attached to each video clip. The assessments (WP and VP) were administered before and after the experimental treatment. Each subject was administered the two assessments consecutive to each other. Since the questions of the modes of assessment were close to identical, it was not appropriate for subjects to communicate between one assessment and the other as this would have contaminated the results. Subjects were required to get on with the second assessment immediately after the first one and were requested not to communicate with each other. The administration of both assessments took approximately one hour 15 minutes at one stretch. Requiring all the subjects to undergo the two assessment instruments irrespective of the teaching environments reduced the threat to external validity of the study as subjects would have been sensitised to the experimental treatment (Cohen and Manion, 1994, pg. 172).

The marking of the athletics pre and post-test scripts was done by trained markers. This was necessary because one of the authors had been engaged in the implementation of the ETUs. This was done to reduce threats to the internal validity of the study by expecting or anticipating that some subjects or a particular teaching environment contributing to better performance (Thomas and Nelson, 1996). Another potential threat of external validity was that the researcher-marker might have being sensitised or sympathetic to the experimental teaching environments (Cohen and Manion, 1994). The markers were final year (fourth year) B.Ed. students, who had PE as one of their two main subjects. The markers had done pedagogic and content courses on athletics. Before the marking exercise, training was given to the markers to explain and instruct them to apply the marking criteria of the two marking schemes of the WP and the VP. The training involved a series of exercises over two sessions of

one and half-hours each. Each marker marked four scripts for each of the 49 subjects. The four scripts comprised two pre-tests and two post-tests.

After all the marking was finished it was decided to select the most reliable marker. This was achieved by calculating measures of inter-marker reliability (Siedentop, 1991) for a series of randomly selected number of post-test scripts marked (72 sub-scores). When the markers' scores were compared, there emerged one marker who scored consistently high on inter-marker reliability when compared with three other markers (85%, 90% and 94%). The marking of this athletic marker was used for the purpose of analysing data.

The efficiency of the three experimental teaching conditions/environments was evaluated by means of paired sample t-tests while correlation was used to measure the relationship between the pre-test and post-test performance of the subjects. AVOVA (analysis of variance) was used to investigate any significant differences between the experimental teaching conditions and the results of the post-test WP and VP for the content domains. MANOVA (multiple analysis of variance repeated measures) was employed to investigate any main effects of independent variables of experimental teaching conditions, gender and attendance on the assessment of content domains. Mean values will be reported in raw scores to facilitate the interpretation of the results and discussion of the findings. Rasch Model analysis was used to measure the unidimensionality and reliability of the WP and VP (Wright & Masters, 1982). The Rasch model is routinely used in Education to generate a single measure of ability for the subjects and a single measure of difficulty for the items of the tests. It converts the raw scores (which are at best ordinal measures) to logits. This unit of measurement (the logit) is a logistic transformation of the raw scores and provides a linear measure of subjects' ability and items' difficulty. Statistically, the appropriateness of the use of the Rasch model was measured using the Infit Mean Square statistic (Wright and Stone, 1979).

Some of the results of this research are reported using the Rasch metric system (logits) and some of the results are reported using raw scores. This was considered to be necessary because not all the audience of the research is familiar with the Rasch models. By keeping this dual format of reporting the results, we ensure that the results of the research will be communicated more easily to a wider spectrum of the readers of the paper.

The statistically significant findings that resulted will be discussed and used as evidence to justify the proposals with regards to the suitability of the modes of assessment and teaching in relation to specific content domains.

Results and Discussion

The validation of the content classification of the items (technique, planning tactics and rules) reached undisputed consensus for all test items of athletics. However, the tri-partite classification of cognitive domains (knowledge and comprehension, application and analysis, synthesis and evaluation) was problematic as each test item was often classified as belonging to more than one of the three cognitive streams

according to the experts. This led to a decision to analyse content domains only (rules, planning tactics and techniques) for the purpose of this discussion.

A Partial Credit Rasch model was used to analyse the fit of the data. The average Infit Mean Square statistic for the WP items was 0.99 (SD=0.33) and for the VP was 0.96 (SD=0.27) which indicates that, overall, the model had good fit on the data (values close to 1 indicate good fit). All items for the WP but one had Infit Mean Square statistic smaller than or equal to 1.3. Rules of thumb suggest that items with Infit Mean Square up to 1.3 are considered to have satisfactory fit for empirical datasets (Bond & Fox, 2001; Karabatsos, 2000). The large majority of the items (32 out of 35) had fit statistics smaller than 1.2. Item 5.1 (Question: What procedure should the runner follow a. towards the end of a sprint?) was the one with the worse fit and had an Infit Mean Square statistic of 1.50. One item from the VP (Item 3 Question: A sprinter changes from 100 meters to 200 meters. Give one adjustment which should be made to the starting position) had a large fit statistic (Infit Mean Square was approximately 1.91). For both tests, the items with large fit statistics were investigated in more depth. However, after further inspection, it was found that the misfit to these items was caused by a very small number of unexpected responses. Because of the small sample size, a small number of unexpected responses may artificially 'inflate' the value of the Infit Mean Square. Overall, it was concluded that the items were of high quality.

The fact, however, that the same items had significantly different fit statistics or difficulty indices on the WP and VP underlines the large impact of the means of the delivery of the tests. The items behaved somewhat differently in the WP compared to the VP. For example, in the VP test, Item 3 was the one with the worse fit because a few subjects with lower total scores scored higher on this item than a few subjects with higher total scores. This, however, was less obvious for the WP version of the item.

Although this phenomenon is not worrying because it happened to a very small extent and to a very small number of items, it indicates the degree to which even small changes to the wording of an item or changes to the administration means of an item can affect its psychometric properties.

There were no severely misfitting response patterns since most of the subject fit statistics were smaller than the 1.3 rule-of-thumb. All the response patterns with more extreme fit statistics were thoroughly scrutinized in order to identify the source of the misfit. It was identified that the source of the misfit was a very small number of unexpected responses to some items. Overall, it was verified that the number of examinees with aberrant response patterns was very small.

The small sample size may have had an impact on the overall fit of the model. Because the number of subjects taking the tests was small, a small number of aberrant responses could introduce too much 'noise' in the data resulting in large fit statistics. Overall, the fit of the data to the model was considered to be satisfactory for all practical intents and purposes of this study.

The satisfactory fit of the data to the model suggests that the items for each assessment mode (WP and VP) form a unidimensional and coherent scale. This

implies that the use of the table of specifications procedure for test construction and the use of validation exercises on the objectives, content and assessment instruments using checklists is a very important process for teachers and paper setters to adopt to improve reliability of assessments.

Since the Rasch model has a satisfactory fit on the data, the properties of unidimensionality and local independence must generally hold. This means that the use of one single score on each test (e.g., the use of subject ability measured in logits) is a good indicator of their overall ability (i.e., it is not necessary to report ability on sub-domains if this is not desirable). Since the property of local independence seems to hold, exposure of the subjects to one item (responding to one item) does not affect significantly his/her probability to respond to another (say, the next) item. This also means that the order by which the items are administered may not have practically significant effect on the responses of the subjects. Overall, the tests appear to have very desirable psychometric properties.

The tests had high reliability indices. The WP reliability was .92 and 0.90 when the pre-test and the post-test data were considered separately. The VP reliability was .92 and 0.86 when the pre-test and the post-test data were considered separately.

The ability for each subject was estimated by the Rasch model (the estimated ability for the subjects is measured in logits). Theoretically, the ability estimates could range from minus infinite to plus infinite. Smaller values indicate smaller ability. The following tables give the distributional properties of the ability estimates of the examinees (in logits).

Table 2: The distributional properties of the ability of subjects

Ability (in Rasch logits)	**pre-test WP**	**post-test WP**	**pre-test VP**	**post-test VP**
Mean	-2.06	-0.82	-1.59	-0.82
Std. Deviation	1.10	0.48	0.68	0.75
Minimum	-3.97	-1.94	-3.35	-2.45
Maximum	-.53	0.46	-0.55	1.25

Comparisons were done between pre- and post-test scores. It was shown that in both the WP and VP the performance of the subjects was increased substantially on the post-test.

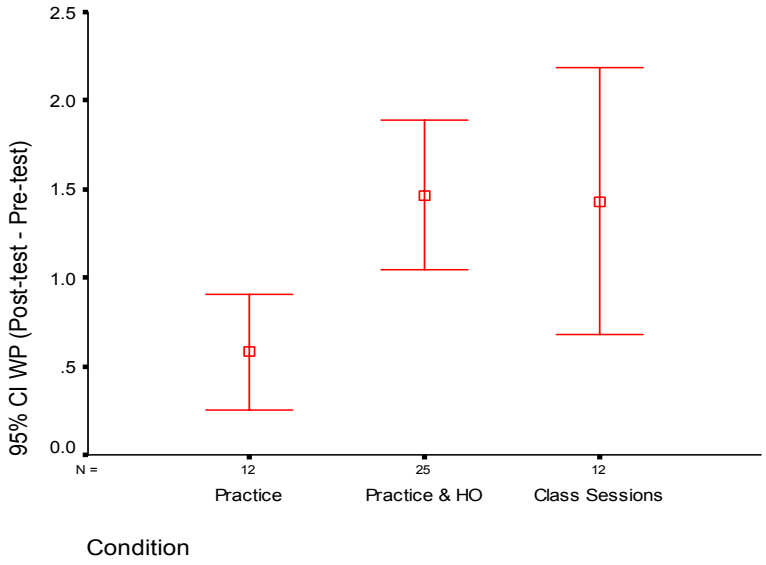
Table 3: Paired performance differences (in Rash logits)

	Mean	**S.D.**	**S.E.**	**95% Conf. Int.**		**t**	**df**	**p**
WP: Pre-test – post-test	-1.24	1.02	0.15	-1.53	-0.95	-8.47	48	<0.001
VP: Pre-test – post-test	-0.77	0.83	0.12	-1.01	-0.54	-6.50	48	<0.001

The effect of the three different teaching conditions/environments on the learning (post-test ability estimate – pre-test ability estimate) of the subjects was investigated.

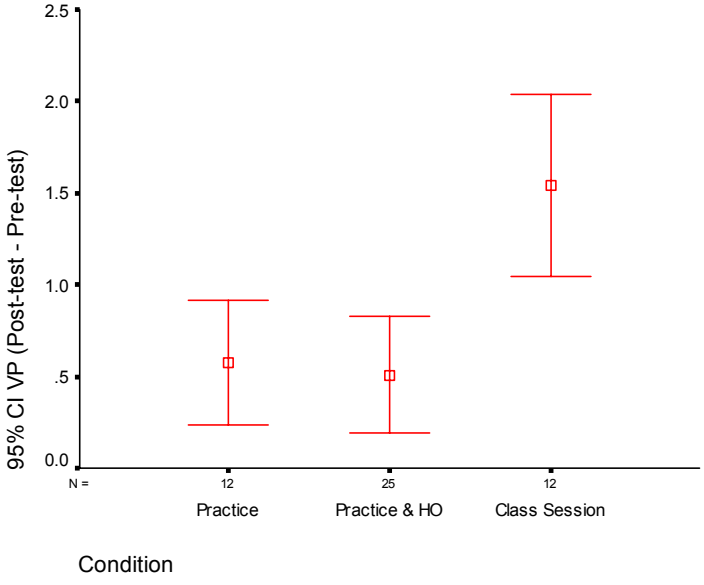
In figures 1 and 2, positive values on the y-axis indicate higher performance on the post-test than on the pre-test. Both figures suggest that all three teaching conditions caused positive changes to the average performance of the subjects. They also indicate that the 'Practice' condition had less effect than the 'Practice & HO' condition and the subjects improved less on the WP. However, no statistically significant difference is shown between the 'Practice' and the 'Class sessions' or between the 'Practice & HO' and the 'Class sessions' on the WP.

Figure 1: Improvement in WP performance by Condition



The 'Class Session' condition had higher impact on the performance of the subjects on the VP than the other two conditions.

Figure 2: Improvement in VP performance by Condition



The three content domains (techniques, planning tactics and rules) and the total scores of the two assessment instruments at post-test will be discussed using ANOVA. This discussion will attempt to illustrate the main effects of the ETUs on the different modes of assessment. The results are presented in raw scores. This information is given in raw scores per sub-domain instead of Rasch logits.

Table 4: Athletics: Experimental Teaching conditions and Mean Scores and S.D of content

Athletics	***Range***	***ETU 1 PE-Athletics class N=12***				***ETU 2 PE-Athletics coursework class N=25***				***ETU 3 Sports Studies Athletics N=12***			
		WP		VP		WP		VP		WP		VP	
Content		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Techniques	0-40	8.23*	4.57	7.75*	6.45	5.28*	6.48	5.14*	4.85	14.50*	5.35	20.67*	8.94
Tactics	0-30	12.17*	6.62	13.25*	7.26	10.60*	4.44	8.36*	4.73	15.75*	5.74	19.00*	6.37
Rules	0-30	13.33*	5.01	13.67*	3.75	10.32*	5.03	7.66*	4.60	23.79*	5.23	24.21*	3.58
Total Score	0-100	33.73*	12.81	34.67*	12.25	26.20*	11.73	21.16*	11.45	54.04*	11.20	64.08*	16.18

WP unseen written paper; VP video based unseen written paper
** Statistically significant at the $p<0.05$ level.*

The differences between the mean scores of the three ETUs were all found to be statistically significant *($p<0.05$*) for the WP and VP. Further investigation of where the differences between the groups lay revealed that for the WP differences emerged between ETU 3 and the other two groups (ETU 1 and ETU 2) at the 0.05 level (Scheffe Multiple Range test) for techniques and rules. For tactics, the significant difference emerged between ETU 3 and ETU 2. This means that the class based teaching sessions yielded significantly better scores for all content domains when the WP was used, however, the scores where less then those achieved on the VP.

For the VP the differences of the mean values were found to be significantly different at the 0.05 level (Scheffe Multiple Range test) between ETU 3 and the other two groups (ETU 1 and ETU 2) and ETUs 1 and 2 for rules. ETUs 1 and 3 and ETU 2 and 3 were found to be significantly different for techniques. Tactics were found to be significantly different between ETUs 2 and 3 only. These findings indicate that class based teaching sessions also resulted in significantly better scores for the all content domains when the VP was administered.

From table 4 it is evident that the mean values of the total score of the subjects in ETU 3, the Sport Studies - athletics (class-based sessions) scored higher than the subjects in the other two groups. Subjects gained higher scores on the VP and the differences of the scores were statistically significant ($F(2,49)$=44.769, p<0.001). The differences of the mean scores of the VP were found to be significantly different amongst ETU 3 and the other two groups at the 0.05 level (Scheffe Multiple Range

test). The differences of the mean scores on the WP were also found to be significantly different ($F(2,49)$=22.339, p<0.001).

Multiple analysis of variance (MANOVA) yielded statistically significant main effects (p<0.05) on experimental teaching environments, gender and attendance with reference to particular content domains of the WP (see table 5). In the case of the VP, the experimental teaching environments emerged as a statistically significant factor (p<0.05) for all content domains. There were no first, second or higher order interactions.

Table 5: Summary of the statistically significant MANOVA results for the athletics WP and VP (n=49) significant at the 0.05 level

Dependent Variables	**Main Effects**	**First order interaction**	**Second order interactions**	**Higher order interactions**
WP-Content				
Techniques	Gender	----	----	----
Tactics	Attendance	----	----	----
Rules	Exp. Teaching Condition.	----	----	----
VP-Content				
Techniques	Exp. Teaching Condition.	----	----	----
Tactics	Exp. Teaching Condition.	----	----	----
Rules	Exp. Teaching Condition.	----	----	----

MANOVA statistically significant main effects can be explained by the substantial increase in the score from the first (pre-test) to the second testing occasion (post-test). In the case of techniques, the statistically significant main effects on gender (p=0.010) for the WP can be explained by the fact that over time, females increased their score more than males on the questions of techniques (see table 6).

Table 6: Athletics: Means and Standard Deviation scores for techniques and gender

Athletics Content: techniques	***Males***	***Females***
Unseen written paper (WP)	***M (S.D.)***	***M (S.D.)***
Occasion 1 (pre-test)	2.03 (3.66)	3.69 (3.74)
Occasion 2 (post-test)	6.80 (6.06)	11.23(7.53)

Although female subjects scored significantly different than males on the techniques component of the WP, both male and female subjects scored higher on the VP. This means that in the absence of a VP, females will do better than males on a WP. Thus, the use of VP for the assessment of techniques is likely to yield higher scores for both males and females, narrowing any inequalities that result in discrimination against male subjects as a result of the use of a WP. However, it remains to be seen (through further investigation) whether the two assessments (WP and VP) actually measure the

same thing. It is possible that the two tests require different skills and may draw on slightly different constructs (knowledge). In that case, although the VP decreases the difference between males and females, the VP may help to reveal significant meaningful differences. The question is: do the two tests measure slightly different things? If yes, then what is the difference? Is it important? If not, let us use the test that does not disadvantage the males unnecessarily. If what the two tests measure is different but informative, then both tests can be used for assessing specific areas of content.

From the evaluation of the written answers given by the subjects for the two modes of assessment (WP and VP), it was observed that some candidates answered differently for the same question. For example, a candidate from PE Athletics class (practice-based class) gave no answer to the question "Give two points of good technique for each of these two methods of baton exchange" in the case of the WP. However, for the very same question in the V.P, this candidate scored marks for giving the correct answers. This finding indicates that the same items had different difficulties on the two modes of assessment. This study does not give explanations to this phenomenon. It would be interesting to further investigate the order of administration of the tests. Further studies need to investigate the use of written and video based assessment in relation to the metal processes and cognitive styles of learners.

Attendance emerged as a statistically significant factor for the WP ($p<0.005$) for the planning tactics component. Subjects who attended for the entire sessions scored higher than those who missed one session. (Occasion 2: Attended all session: $m=12.82$ S.D=5.83; Missed one session $m=10.93$ S.D.=5.17). Attendance did not result as a significant factor for the content domains of rules and techniques, possibly because these were consolidated regularly during the sessions, whereas planning tactics was more related to a specific activity (plan a 100m race: session 2; plan a 200m race: session 3) presented in each particular session.

For techniques assessed through the VP there emerged a statistically significant main effect on the experimental teaching environments ($p<0.01$). This is attributed to the fact that at post-test level subjects in the Sports Studies-Athletics class (class-based sessions) outperformed the subjects in the other two groups for the VP. The improvements between pre and post-test phases can be seen in table 7. The same result emerged in the case of planning tactics in the VP ($p<0.05$). This implies that when the assessment mode was the VP, the practice-based classes did not obtain superior skills of knowledge recall, application and evaluation of tactics through listening, watching and performing demonstrations or practising and reading about planning tactics. On the other hand the class-based group received visual, verbal and written engagement with knowledge on tactics, and this resulted in better learning outcomes.

Table 7: Athletics: Means and Standard Deviation scores for content and experimental teaching environments.

Athletics Video based unseen written paper (VP)	***ETU 1 PE-Athletics class N=12***	***ETU 2 PE-Athletics coursework class N=25***	***ETU 3 Sports Studies Athletics N=12***
	M (S.D.)	***M (S.D.)***	***M (S.D.)***
Techniques Occasion 1 (pre-test)	4.00 (4.18)	1.48 (2.57)	4.92 (4.03)
Techniques Occasion 2 (post-test)	7.75 (6.45)	5.14 (4.85)	20.67 (8.94)
Planning Tactics Occasion 1 (pre-test)	9.42 (6.63)	6.56 (4.78)	6.75 (3.77)
Planning Tactics Occasion 2 (post-test)	13.25 (7.25)	8.36 (4.73)	19.00 (6.37)

Table 8: Athletics: Means and Standard Deviation scores for rules and experimental teaching environments.

Athletics Content: rules Unseen written paper (WP)	***ETU 1 PE-Athletics class N=12***	***ETU 2 PE-Athletics coursework class N=25***	***ETU 3 Sports Studies Athletics N=12***
	M (S.D.)	***M (S.D.)***	***M (S.D.)***
Occasion 1 (pre-test)	7.67 (5.65)	2.04 (2.55)	3.75 (6.15)
Occasion 2 (post-test)	13.33 (5.01)	10.32 (5.03)	23.79 (5.23)
Video based unseen written paper (VP)			
Occasion 1 (pre-test)	5.67 (4.93)	4.38 (5.20)	6.87 (5.17)
Occasion 2 (post-test)	13.67 (3.75)	7.66 (4.60)	24.21 (3.58)

The experimental teaching environments resulted in statistically significant main effect on the rules content of both the WP ($p<0.001$) and the VP ($p<0.001$) (see table 8). The subjects in the Sports Studies-Athletics group (class-based sessions) outperformed the subjects in the other two groups. It is observed that the practice and the class-based groups scored lower on the WP. Thus, one can propose that for athletic type activities, rules questions are better assessed using the VP. One can also suggest that teaching rules in a class-based environment would result in higher scores whether written or video based assessments are used.

The classroom environment was a predominant factor in this study. The use of assessments for the "class-based sessions" and the more traditional practice-based PE classes was somewhat new to research on teaching in physical education.

Conclusions

The study was aimed at establishing a process of constructing and validating ETUs and assessment instruments that contribute to higher reliability of assessments. The process of construction and validation of coursework objectives, content and

assessment instruments involved using established procedures and statistical analysis. The development of objectives and content (Linn and Grondlund, 1995); the use of cognitive taxonomies of education (Bloom, 1956, cited in Linn and Grondlund, 1995) and the application of the 'table of specifications' (Ebel and Frisbie, 1991) contributed to a sound frame work. The use of criteria for the validation by expert-teachers (Vickers, 1990) included appropriateness, technical soundness, completeness and feasibility for the instructional and assessment objectives and content. Relevance, clarity, correctness and technical soundness were used for the marking scheme (questions and answers). Appropriateness and technical soundness was the criteria for the validation of the video clips with reference to the adjoining questions.

The use of the Rasch model provided a very strong context for the evaluation of the tests and gave insights to the psychometric characteristics of the test items and the responding behaviour of the subjects. The Rasch analysis indicated that the thorough process for the development of the tests paid off because it led to reliable and unidimensional measurement instruments.

The MANOVA (repeated measures) analysis provided the details of the statistically significant improvement shown by the various different teaching environments over the treatment period. The findings of this study suggest that the teaching of content (i.e. techniques, planning tactics and rules) in a practical performance teaching situation does not prepare students for video based assessment. These results of this study further suggest that the teaching environment (i.e. performance-based vs. class-based) is sensitive video-based assessments for all the content domains of techniques, planning tactics and rules in the case of athletics. From this study it can be concluded that from the two constructed and validated modes of assessments the VP was sensitive to teaching environments for all content domains.

The results of this study indicate that when assessment was done using the written paper, the teaching environment (i.e. performance-based vs. class-based) was sensitive to the content domain of rules only, and not to techniques and tactics. However, the scores for rules where significantly higher when the assessment was video based. This implies that when the rules content of athletics is taught in a class based sessions, students will get significantly better scores when assessed with video based written papers. However, given that secondary PE programmes include various families of activities such as athletics, dance, games, gymnastics, outdoor pursuits and swimming, inferences and generalisations of the findings of this study cannot be made until further investigation on the teaching and assessment of these activities is carried out.

The findings of this study have implications for teaching and assessments for teachers and teacher educators. Specific content domains such as techniques, tactics and rules should be treated with more focus on the learning pedagogy and variety of assessment instruments. The research outcomes of this study challenge dominant learning pedagogies and assessment practices of physical education.

References

AEB (1996) *GCE Advanced Level Physical Education and Sport Studies: Examiners Report*, Surrey: Associated Examining Board

AEB (1997) *GCE Advanced Level Physical Education and Sport Studies Examiners Repor,* Surrey: Associated Examining Board

AENA (1994) *Netball: Know the game series,* London: A&C Black

AENA (1996) *Netball: Students information pack,* Hertfordshire: AENA

AENA (1997) *Official rules of the International Federation of Netball Associations (IFNA),* Hertfordshire: AENA

Alderson, J (1988) Examination syllabuses and curriculum development in PE, *British Journal of Physical Education,* 19(6), pp. 214-216.

Armstrong, N (1996) *New Directions in Physical Education,* London: Cassell

Aylett, S (1990) Is GCSE the cross-roads for Physical Education? *British Journal of Physical Education*, 21(3), pp. 333-337.

Beashel, P Sibson A, Taylor, J (1997) *The World of Sport examined-CD-Rom,* Surrey: Nelson Thornes

Bloom, BS (1956) *Taxonomy of educational objectives: The classification of educational goals. Handbook 1. Cognitive domain*, New York: McKay

Bond, T, & Fox, C (2001) *Applying the Rasch model: fundamental measurement in the human sciences*, New Jersey: Lawrence Erlbaum Associates

British Athletic Federation (BAF) (1994) *Track athletics: Know the game series,* London: A&C Black

Buck, MM, Harrison, JM, and Bryce, GR (1991) An analysis of learning trails and their relationship to achievement in volleyball, *Journal of Teaching Physical Education,* 10*,* pp. 134-152.

Carr, GA (1999) *Fundamentals of track and filed(2nd Ed),* Leeds: Human Kinetics

Carroll, B (1990a) Examinations and assessments in Physical Education. In N Armstrong (Ed) *New Directions in Physical Education, Vol 1.* Leeds: Human Kinetics.

Carroll, B (1990b) The twain shall meet: GCSE and the National Curriculum, *British Journal of Physical Education*, 21(3), pp. 329-332.

Cassar, G (1995) Assessments and examinations in Physical Education, Unpublished M. Ed. Dissertation, Manchester: University of Manchester

Carroll, B (1994) *Assessments in Physical Education*, London: The Falmer Press

Chung, TW (1992) The effectiveness of computer based interactive video instruction on psychomotor skill analysis competency of pre-service PE teachers in tennis teaching, Unpublished doctoral dissertation (EDD) University of Northern Colorado

Ebel, RL (1965) *Measuring educational achievement*, Englewood Cliffs, NJ: Prentice Hall

Ebel, RL and Frisbie, DA (1991) *Essentials of educational measurement (5th Ed),* Englewood Cliffs, NJ: Prentice hall

Eckrich, J, Widule, CJ, Shrader, RA, and Maver, J (1994) The effects of video observational training on video and live observational proficiency, *Journal of Teaching in Physical Education,* 13*,* pp. 216-227.

Fleming, S (1994) Understanding 'understanding': Making sense of the cognitive approach to the teaching of games, *Physical Education Review,* 17(2), pp. 90-96.

Forsyth, S (1994) Thoughts on the first year of Higher Grade Physical Education, *Scottish Journal of Physical Education,* 22(3), pp. 5-21.

Francis, JM (1990) 'A' level examinations in Physical Education and Sports Studies: Open to all, *British Journal of Physical Education*, 21(3), pp. 338-340.

French, KE, and Thomas, JR (1987) The relation of knowledge development to children's basketball performance, *Journal of Sport Psychology, 9,* pp. 15-32

Harrison, JM, Preece, LA, Blakemore, CL, Richards, RP, Wilkonson, C, and Fellingham, GW (1999) Effects of two instructional models - Skill Teaching and Mastery Learning - on skill development, knowledge, self-efficacy, and game play in volleyball, *Journal of Teaching in Physical Education,* 19(1), pp. 34-57.

Hoffman, SJ (1974)Taking the fun out of skill analysis, *Journal of Teaching Physical Education, Recreation and Dance,* 45(9), pp. 74-76.

Honeybourne, J, Hill, M, Moors, H, (2001) *Advanced Physical Education and Sport for A Level (2nd Ed),* Cheltenham,UK: Nelson Thornes

Honeybourne, J, Hill, M, Moors, H, (2002) *Advanced Physical Education and Sport Teacher Resource pack (2nd Ed)* Cheltenham,UK: Nelson Thornes

IAAF (1990) *IAAF Level 1: Techniques of athletics and teaching progressions,* Monaco: IAAF

Ignico, A (1994) A comparison of videotape and teacher-directed instruction on knowledge, performance and assessment of fundamental motor skills*, Journal of Educational Technology Systems,* 23 (4), pp. 363-368.

Jewett, AE, Bain, LL, and Ennis, CD (1995) *The curriculum process in Physical Education (2nd Ed)* Dubuque, IA: Brown and Benchmark

Karabatsos, G (2000) *A critique of Rasch residual fit statistics*, Journal of Applied Measurement, 1, pp. 152-176.

Lee, AM (1997) Contributions on student thinking in Physical Education, *Journal of Teaching Physical Education*, 16*,* pp. 262-277.

Linn, RL and Grondlund, NE (1995) *Measurement and assessment in teaching (7th Ed)* New Jersey: Merrill/Prentice Hall

MEG (1993) *GCSE Physical Education Chief Examiners Report,* Cambridge: Midland Examining Group

Melograno, VJ (1996) *Designing the Physical Education curriculum (3rd Ed)* Champaign, IL: Human Kinetics

Morrison, C and Reeve, J (1988) Effect of instruction and undergraduate major on qualitative skill analysis, *Journal of Human Movement Studies*, 15, pp. 291-297.

Pinheiro, VED and Simon, HA (1992*)* An operational model for motor skill diagnosis, *Journal of Teaching in Physical Education,* 11, pp. 288-302.

Piotrowski , S and Capel, S (2000) Formal and informal models of assessment in physical education. In S Capel and S Piotrowski (Eds), *Issues in Physical Education.* London: Routledge-Falmer.

Pocock, T (1995) *Official rules of sports & games 1995-96 (19th Ed)* London: Hamlyn

Roscoe, J (1996) *Physical education Advanced level-Athletics (3rd Ed)* Cheshire: Jan Roscoe Publications

Roscoe, J (1998) *Physical Education Publishing Catalogue,* Cheshire: Jan Roscoe Publications

Rink, JE, Werner, PH, Hohn, RC, Ward, DS, and Timmermans, HM (1986) Differential effects of three teachers over a unit of instruction, *Research Quarterly for Exercise and Sport,* 57(2), pp. 132-138

Scott, T (1999) *GCSE PE for Edexcel Teacher's Pack (2nd Ed)* Oxford: Heinmann

Scott, T (2001) *GCSE PE for Edexcel (2nd Ed)* Oxford: Heinmann

Siedentop, D, Doutis, P, Tsangaridou, N, Ward, P, and Rauschenback, J (1994) Don't sweat gym! An analysis of curriculum instruction, *Journal of Teaching in Physical Education,* 13, pp. 375-394.

Silverman, SJ (1997) Technology and Physical Education: Present possibilities, and potential problems, *Quest,* 49, pp. 306-314.

Silverman, SJ, Subramaniam, PR, and Woods, AM (1998) Task structures, student practice and skill in Physical Education, *Journal of Educational Research,* 91*,* pp. 298-306.

Stirling, S and Scott, I (1989) The knowledge and understanding assignment in Standard Grade Physical Education, *Scottish Journal of Physical Education*, 17, pp. 3-13.

Sweeting, T and Rink, J (1999) Effects of direct instruction and environmentally designed instruction on the process and product characteristics of a fundamental skill, *Journal of Teaching Physical Education,* 18, pp. 216-233.

Thorpe, RD and Bunker, DJ (1986) Landmarks on our way to 'Teaching for understanding'. In RD Thorpe, DJ Bunker, and L Almond (Eds), *Rethinking games teaching*. Loughborough: Loughborough University of Technology.

Thorpe, RD, Bunker, DJ, and Almond, L (1986) A change in focus for the teaching of games. In M Piéron and G Graham (Eds), *Sport pedagogy: The 1984 Olympic Scientific Congress Proceedings, Vol 6.* Champaign: Human Kinetics.

ULEAC (1997) *GCSE Physical Education Examiners' Comments,* London: University of London Examinations and Assessment Council

Vickers, JN (1990) *Instructional design for teaching physical activities,* Champaign,

Walker, LR (1987) Athletics. In R McGee and A Farrow (Eds), *Test questions for physical activities.* Champaign, IL: Human Kinetics.

Wilkinson, S (1991) A training programme for improving undergraduates' analytic skill in volleyball, *Journal of Teaching in Physical Education,* 11, pp. 177-194.

Wright, B D, & Masters, G N (1982) *Rating scale analysis*, Chicago: MESA Press

Wright, B D, & Stone, M H (1979) *Best test design*, Chicago: MESA Press