IRT 在体育学习成果测评领域中的应用及其对我国体育中考的启示

何毅，董国永
（华中师范大学 体育学院，湖北 武汉 430079）

摘 要：对项目反应理论(IRT)在美国体育学习成果测评体系(PE Metrics)中的应用进行解析，并提炼了基于 IRT 的 PE Metrics 测评体系表现特征。研究认为，IRT 在 PE Metrics 中的应用主要体现在基于测验等值技术的数据采集设计、利用多层面 Rasch 模型进行参数估计、通过标定与校准建立体育学习成果测评题(项目)库等方面。在 IRT 方法和技术的支持下，PE Metrics 具备测评项目及量规的难度参数恒定且分布均匀、不同运动项目的测评结果可进行交流、学生能力估计的精确性、测评项目开发的动态性和可延续性等特征。在此基础上，提出完善我国体育中考测评体系的应然之策，主要包括：广泛纳入“真实性”运动技能测评内容，突破体育中考的“应试化”桎梏；利用测验等值技术，实现体育中考运动技能测评分数的可比性；研制参数详实的测评工具，提高体育中考分数的精确性和区分度；建立动态体育中考题(项目)库，不断丰富和完善体育中考测试内容。

关 键 词：学校体育；体育学习成果测评；体育中考；项目反应理论
中图分类号：G807 **文献标志码：**A **文章编号：**1006-7116(2021)04-0094-07

Application of IRT in the field of physical education learning achievement evaluation and its enlightenment to China's physical education examination for high school

HE Yi，DONG Guo-yong
(School of Physical Education，Central China Normal University，Wuhan 430079，China)

Abstract: This paper analyzes the application of IRT in PE metrics of American sports learning achievement evaluation system, and refines the performance characteristics of PE metrics evaluation system based on IRT. The research shows that the application of IRT in PE metrics is mainly reflected in the design of data collection based on test equivalence technology, parameter estimation by using multi-level Rasch model, and the establishment of test (item) database of physical education learning achievements through calibration and correction. With the support of IRT method and technology, PE metrics has the characteristics as follows: constant and uniform distribution of difficulty parameters of evaluation items and rubrics, communication of evaluation results with different sports items, accuracy of students' ability estimation, and dynamic and continuity of development of evaluation items. On this basis, this paper puts forward the corresponding measures to improve the evaluation system of China's physical education examination for high school, which mainly includes: widely introducing the "authenticity" sports skills evaluation content, breaking through the "exam oriented" shackles of physical education examination for high school; realizing the comparability of sports skills evaluation scores by using test equivalent technology; developing

收稿日期：2020-12-11
基金项目：国家社会科学基金项目(19CTY008)；湖北省高校省级教学研究项目(2017082)；中央高校基本科研业务费资助(优博培育项目)(2020YBZZ066)。
作者简介：何毅(1994–)，男，博士研究生，研究方向：学校体育。E-mail：66706547@qq.com 通信作者：董国永

the test instruments with detailed parameters in order to improve the accuracy and differential degree for the scores of physical education examination for high school, and establishing a dynamic database of physical education examination programs (items) to constantly enrich and perfect the contents of the physical education examination.

Key words: school physical education；physical education learning achievement evaluation；physical education examination for high school；item response theory

2020 年 8 月体育总局与教育部联合印发的《关于深化体教融合 促进青少年健康发展的意见》和 2020 年 10 月中共中央办公厅、国务院办公厅印发的《关于全面加强和改进新时代学校体育工作的意见》提出：“将体育科目纳入初、高中学业水平考试范围。改进中考体育测试内容、方式和计分办法，科学确定并逐步提高分值。”[1-2]由此可见，随着体育在学校教育中地位不断提升，学生的体育考试成绩将不再是无关紧要的分数，而是更有可能成为学生综合素质评定甚至是中高考的重要组成部分。与此同时，在体育考试逐渐演变为“高利害”测试的背景下，体育考试分数必然会成为学生、学校和社会关注的焦点。因此，如何确保体育“高利害”测试的科学性、合理性和公平性成为决定我国体育考试制度有效实施的关键所在。然而，从现阶段我国体育中考实施效果来看，虽基本满足体育“高利害”测试的制度要求，但仍存在争议，有待进一步完善，如“应试化”倾向严重、评分标准区分度不足、项目设置不尽合理等[3]。究其原因，现阶段我国体育中考的测评方法和技术难以满足体育“高利害”测试的科学性、合理性和公平性需求，成为掣肘体育中考测试内容选择、测评方式改进和计分办法更新的重要因素。

现阶段我国大部分地区体育中考的测量标准和工具均是以经典测量理论(Classical Test Theory，CTT)为基础而构建的，因此难以克服其理论体系的先天局限性，如弱或差的信效度控制、孤立的测试开发、评价结果无法进行交流等[4]。为了克服 CTT 的局限性，一种新兴的测量理论——项目反应理论(Item Response Theory，IRT)逐渐发展起来。基于 IRT 设计的标准化测验不仅在理论上更加符合测量原理，同时也具有更大的解决实际测量问题的潜力，因此在现代心理和教育测量领域得到了广泛应用[5]。然而，在当前我国体育学习成果测评领域，IRT 的研究与实践应用不足，在一定程度上限制了我国体育测量特别是体育中考测评改革与发展的推进步伐。美国国家运动与体育教育协会(National Association for Sport and Physical Education，NASPE)研制推出的体育学习成果测评体系——PE Metrics(简称 PEM)正是以 IRT 为基础，充分利用现代测量理论与方法的优势，突破传统体育学习成果测量的局限性，在体育学习成果测量实践中取得显著效果，为强调问责与改进的美国学校体育作出了突出贡献。实践证明，PEM 不仅获得了体育教师的广泛认可，同时也是体育科研人员较为信赖的大范围体育学习成果测评工具[6]。它山之石，可以攻玉。本研究从研制方法和技术层面深入剖析 IRT 在美国 PEM 测评体系中的应用，总结归纳基于 IRT 的 PEM 的表现特征，并针对我国体育中考所面临的现实困境，提出完善我国体育中考测评的应然之策，为进一步推进我国体育中考改革助益。

1　基于 IRT 的 PEM 研制

PEM 是针对美国 K-12 年级的以标准为参照的体育学习成果测评体系。NASPE 成立的评价工作组(Assessment Task Force，ATF)依据美国国家体育课程标准(以 2013 年版为例)的 5 个领域目标，开发涵盖了两个维度的评价内容，即针对标准 1 的运动技能评价和针对标准 2~5 的认知评价。对于标准 1 统领的运动技能评价，ATF 根据不同评价(运动)项目或任务制定了详细的评价量表，每份量表中均包含有表现性指标、评价任务、评分量规和评价方案、设备或材料、空间或位置图等内容，评价者依据评分量规对学生在评价任务中的表现进行打分，而评价方案、设备或材料、空间或位置图等则主要用于评价过程中对学生和评价者的详细指导；对于标准 2~5 所涵盖的知识、概念和态度，因为难于将其操作化，ATF 最终决定采用纸笔测验形式对学生进行考核[7]。此外，ATF 还开发了网络在线平台 PEM 在线(PE Metrics online)，其主要功能包括：为评价者提供更加直观的视频指导，帮助评价者提高评分准确性和操作熟练程度；提供已开发或后续开发的评价工具；帮助评价者录入、分析和解释评价结果，为评价的利益相关者提供反馈信息[8]。

1.1　IRT—PEM 研制的理论基础

CTT 作为历史上第一个测验理论，经过多年探索与发展已经形成了一套较为完整的理论体系，是过去测量实践中使用较为广泛的理论模型。然而，CTT 也存在着无法克服的先天缺陷，包括无法区分各类测量误差、样本依赖性、能力量表与难度量表的不一致性等[9]。20 世纪 50 年代，在分析和克服 CTT 自身不足和

缺陷的基础上，加之电子计算机的普及与发展，一个更加复杂、统计效率更高的测量理论模型——IRT 应运而生。

IRT 的主要内容是通过数学函数揭示被试者在测验项目上的反应行为与被试者潜在特质之间的关系。这种关系函数表达式，即项目特征曲线解析式，被称为 IRT 各种模型的项目反应函数[10]。常用的 IRT 模型有正态卵形模型，单、双、三参数 Logistic 模型，其中单参数 Logistic 模型也被称 Rasch 模型。在实际应用中，通过这些模型对测验分数进行统计调整，能有效解决测量实践中测验分数等值、项目参数估计和误差控制等问题。随着 IRT 模型的不断丰富和扩展，其逐渐实现了对人格特质、潜在能力、行为意向、情景评价等多种目标的测量。如今，IRT 已成为教育领域几项重要测验的基石，如美国研究生入学考试(GRE)、学术评估测试(SAT)以及中国大学生英语水平测试(CET)等。

IRT 的测量优势主要体现在以下几个方面：第一，题目参数的不变性。IRT 的题目参数估计是独立于考生样本的，即题目难度不会因为抽样学生能力水平的高低而变化；第二，题目参数与能力参数的一致性。项目反应理论将项目难度和被试者能力置于同一尺度，使用共同的 Logit 单位；第三，误差控制的精确性。在测验中，不同能力或得分有其不同的概率误差。项目反应理论通过提供题目信息函数和测验信息函数两个统计量，控制不同能力水平被试者的测量误差，从而更精确地估计每个考生的能力水平[11]。总体而言，IRT 具有诸多 CTT 所不具备的优势，是现代心理和教育测量实践中最受欢迎的测量理论之一。

1.2 IRT 在 PEM 研制过程中的应用

1)基于测验等值技术的数据采集设计。

在教育和心理测量实践中，往往需要通过多种测验形式来测量同一知识结构或心理品质，为了使不同测验形式的分数建立在同一尺度之上，进而比较不同测验形式中受试者的能力水平，就需要对测验分数进行等值处理。因此，测验等值对于测验结果的可比性、保证测验的公平性具有重要意义。当不同测验形式分别施测于不同被试组时，等值需要完成参数量表的变换，即将不同被试群体的参数标刻在同一参数量表之上，而实现变换的前提是不同测验形式必须有公共测验题(项)目相关联，即铆测验–非等组设计。因此，ATF 在全国性数据采集中使用了水平和垂直等值设计，其包含共同项目(试题)和连接项目(试题)。共同项目用于校准同一年级中不同项目，而连接项目用于关联不同年级之间项目。共同项目和连接项目的选择并不是事先预定的，而是根据试点测试阶段数据的常规项目分析结果决定。其中，共同项目是在常规项目分析结果中显示出良好区分度的项目，连接项目则是根据难度水平进行选择，如“原地运球”和“单脚跳”是幼儿园评价中的共同项目，“滑步”和“用拍击球”是用于连接幼儿园和 2 年级之间的连接项目[12]。可以看出，在基于 IRT 的测验等值技术指导下 ATF 制定了科学合理的数据采集方案，为后续项目分析与校准奠定基础。

2)利用多层面 Rasch 模型进行参数估计。

ATF 在不断对测评(项目)进行修改和完善后，利用广泛的项目管理网络从全国各地进行数据采集，以进行后续数据分析和校准。具体而言，ATF 专门雇佣评分人员根据测试录像和评分量规对学生运动表现进行评分[13]。评分数据使用传统的和基于 IRT 的两种方式进行分析。首先采用描述性统计分析，对数据中的异常值或打字错误进行筛选、识别和删除；然后计算项目反应频率，以及每项评价的平均值和标准偏差；最后，使用多层面 Rasch 模型分析不同测评项目、评分量规和学生能力水平。多层面 Rasch 模型是经过拓展的 Rasch 模型之一，其主要作用在于通过被试者在题(项)目上作出特定反应概率来计算个体能力和题(项)目难度。评分数据的多层面 Rasch 分析是通过 FACETS 软件完成的，其报告结果包括项目及量规难度、学生能力水平、残差均方和加权后的残差均方。项目及量规难度值和学生能力水平值均以 logit 为单位，这也实现了题目难度与学生能力水平的参数估计及校准。ATF 的统计分析结果表明，PEM 的采集数据与模型拟合良好，量规及评价项目难度等分布均匀[12]。

3)通过标定与校准建立体育学习成果测评题(项目)库。

在数据分析过程中，ATF 首先对特定年级的评分数据进行分析并锚定，再分析其他年级评价项目的统计数据。如在运动技能评价(项目)构建中，首先分析 2 年级的评分数据，然后在 2 年级的尺度上对其他年级数据进行分析，最终将所有年级的量规及项目都标定在同一尺度之上。事实上，PEM 的构建正是遵循了题库开发的基本程序，在将所有测评题(项)目和量规都置于同一尺度之后，就形成了一个包含不同项目及其相关统计资料(如难度)的体育学习成果测评资源库。从 PEM 的研制流程和成果发布可以看出，题库建设并不是一蹴而就的，而是一个动态的持续不断的过程。如在 PEM 的研制过程中，ATF 首先发布了针对小学阶段的运动技能测评工具[14]，随着测评项目和测试工作的逐步完成，NASPE 又陆续发布了小学和中学阶段的运动技能和认知测评工具[15]，由此逐步构建中小学体育学习成果测评体系 PEM 的基本框架。此后，通过 IRT

的标定和校准，ATF 不断地丰富和完善测评工具与内容，最终形成了一个资源丰富、交互共享的 K–12 年级体育学习成果测评题(项目)库。

2　基于 IRT 的 PEM 表现特征分析

2.1　测评项目及量规的难度参数恒定且分布均匀

在以 CTT 为基础的测量实践中，对于项目难度、区分度等参数的估计是根据测试样本获得的，因此，样本代表性直接影响着参数值的大小。以难度参数而言，对于同一个测验项目，若测试样本的整体水平较高，就会过低地估计项目难度值；若测试样本的整体水平较低，则会过高地估计项目难度值。相反，在 IRT 中难度被认为是题(项)目的固有属性，其估计得出的参数不受样本能力水平的影响，即参数不变性。PEM 正是利用了 IRT 的这一特性，准确估计出评价项目及量规的难度参数值，很好地解决了样本依赖性问题，从而确保评价工具的有效性和可靠性。此外，在 PEM 的构建过程中，ATF 经过多次实地测试和项目分析，并根据分析结果对测评(项目)进行修改，其目的在于确保项目及量规的难度参数适当，即既要保证项目及量规难度范围的广度，也兼顾其难度参数的连续性。在实际应用中，由于测评项目及量规是恒定的并且是已知的，教师或研究人员可以根据评价目的和意图形成测验。例如，若想了解学生的整体能力水平，那么就可选择难度范围较广的测评项目；若想构建标准参照类型的测试(如资格证考试)，则可选择与截至分数(或标准)难度相当的测评项目。

2.2　不同运动项目的评价结果可进行交流

PEM 包含两个维度的测评内容，即针对标准 1 的运动技能测评和针对标准 2~5 的认知测评。其中，标准 1 引领的运动技能测评中包含多个运动项目或任务，彼此之间的内容也大不相同。在 CTT 中真分数的意义仅仅限于一组特定的测评项目，因此无法建立不同运动技能测评结果之间的联系，这也进一步限制了测评项目的丰富和测评结果的应用。相对而言，基于 IRT 的等值技术为这一问题提供了很好的解决途径。在 PEM 的构建过程中，ATF 利用 Rasch 模型进行校准，使所有项目、量规及学生能力置换于同一量表之上，进而使得不同项目之间、不同量规之间、项目及量规与学生能力之间可以进行比较，很好地解决了不同测验版本之间的等值问题。如在 PEM 中，参加篮球运动项目测评的学生成绩可以直接和参加排球运动项目测评的学生成绩进行比较，但前提是需将学生在量规上的得分转换为“能力分数”。

此外，评价分数可进行比较的另外一个好处就是，可以测量学生成绩的变化和增长。如某学生在 2 年级时参加“立定跳远”项目测评，但随着该学生年级升高和教学内容及难度变化，其在 5 年级须参加“体操”项目测评，这种情况下通过 PEM 依然可以比较学生随着年级变化的能力水平。评价结果可进行交流这一特性使得 PEM 具备应用于大规模标准化运动技能测试的潜力，确保了大范围评价分数的统计学意义，这也是 PEM 逐渐开始应用于体育科研领域的重要原因之一。

2.3　学生能力估计的精确性

传统体育学习成果测评方式是以常模参照为基础的，这就意味着只能通过将个人成绩与常模团体进行比较，进而判断个体在团体中的相对位置和名次，但无法准确判断学生学习目标的达成情况。相较而言，标准参照评价更加关注个体对知识和技能掌握的真实情况，是一种以过程性评价为主，过程性评价与终结性评价相结合的评价范式。因此，标准参照评价可以更准确诊断学生的学习成果。在 PEM 构建过程中，ATF 通过解析“课程标准”，撰写具有可操作性的表现性指标和评价量规，进而开发出相应的测评内容和方式，其目的在于构建标准参照的体育学习成果测评体系，即 PEM。此外，在形成测评题(项目)库之后评价可以选择与学生能力相当的评价项目，进而精准定位学生的能力水平。

实际上，为了确保测评结果的精确性和可靠性，ATF 在 PEM 的构建过程中做了大量工作。如在测评项目及量规的参数估计过程中，ATF 通过不断测试、反馈及修订，确保测评项目及量规难度参数的连续性，从而为评价者提供更加精确的学生能力水平信息。此外，为了检验 PEM 测量准确性，ATF 成员还对测验中所需最少的测评项目数量这一问题进行验证。结果表明：当 PEM 用于“高利害”测试时，应当使用至少两个测评项目精准定位学生能力；而在教学实践中，依然可以使用单一测评项目去确定学生的运动水平，只是需要教师更加谨慎地对结果进行解释[16]。

2.4　测评(项目)开发的动态性和可延续性

基于 IRT 的题库建设是现代教育测量领域的主流趋势，并在各个学科领域得到了广泛应用。虽然 PEM 根据课程标准的年级水平划分包含各个年级特有的测评项目，但从本质上来说，在将所有评价项目置于同一尺度之后，年级水平仅仅起到参考作用，而整个测评项目所组成的项目库才是其实质所在。换言之，评价者不一定需要局限于从特定年级的评价项目中选择测评工具，而可以从整个题(项目)库中选取合适测评工具。除此之外，题库建设实现了测评(项目)开发的动态性和可延续性。过去以 CTT 为基础的测量实践中，

因其信效度和误差控制问题，大多测试都是孤立开发的，无法对其进行改进和完善。而在 PEM 中，通过对评价项目及量规进行标定和校准，从而使所有评价项目都置于同一尺度，在后续也可以依照已有项目尺度增添新的测评项目。实际上，从 PEM 的研制流程及成果发布上也可以看出其评价开发的动态性和可延续性特征。由于测评体系研制是一个极其复杂又耗费资源的过程，特别是在需要进行全国性测试和数据采集的情况下，ATF 通过将研制任务阶段化，即在开发小学标准 1 测评项目之后继而开发中学测评项目，逐步实现对测评题(项目)库的构建。

3 对我国体育中考的启示

多年来体育中考为我国学校体育发展带来的积极效应显而易见，而且体育中考经过多年改革与发展，在不断实现自我完善的同时，也为各学段学生综合素质评定中体育评价及体育高考积累丰富和宝贵经验，进一步推进了我国学校体育评价与考试制度的改革与发展步伐。事实上，虽然我国在体育中考改革与发展进程中积累了一定经验，但在体育中考实践中存在一些问题或不足，亟待进一步解决和完善。如体能性、碎片化的考试内容致使体育中考的“应试化”倾向严重；评分标准的科学性、公平性欠缺，体育中考分数的真实性和有效性大打折扣；体育中考与学校体育课程教学缺乏有效衔接，“考什么，教什么”使得课程标准的效力削减等[18]。简言之，体育中考处于“风口浪尖”的重要原因是其考试性质发生了根本性改变，即由水平性考试转变为选拔性考试，而沿用传统测量技术、考核方式、评分标准等内容显然很难满足选拔性考试需要，其结果必然导致体育中考的部分功能难以显现或缺失。因此，在体育中考上升为国家战略且具有法律效应的既定事实下，改进与更新传统体育中考的测评方式、评价内容和计分标准等具体操作，是完善我国体育中考制度、推进体育中考改革与发展的关键。因此，我国体育中考应借鉴美国 PEM 的成功经验，充分利用现代测量理论 IRT 的方法和优势，突破传统体育学习成果测评的现实困境，以建立科学、合理和公平的体育考试制度。

3.1 广泛纳入“真实性”运动技能测评内容，突破体育中考的“应试化”桎梏

2020 年 10 月中共中央、国务院印发的《深化新时代教育评价改革总体方案》指出：“改进中考体育测试内容、方式和计分办法，形成激励学生加强体育锻炼的有效机制。”可以看出，体育中考不仅是对学生体育学习效果进行总结性评价，而是希望成为促使学生养成良好锻炼习惯进而提升体质的长效手段。然而，近年来随着体育科目分值在中考成绩中的比重逐年提升，体育中考中的“异化”和“应试化”现象也越来越严重，关于体育中考中“吃药”“潜规则”“考前突击”等事件的报道屡见不鲜[19]。究其原因，体育中考的项目设置及测试内容不尽合理。如目前各省市中考体育测试项目中技能项目明显少于素质项目，且部分技能项目也仅仅是单个动作考试(如篮球投篮、足球颠球)，忽视体育运动的技能性、情境性特点[20]。单一化和机械性的考试内容使得学生、家长和学校能够通过短时间集训“应对”体育中考，使得正常体育教学秩序也受到严重干扰，成为“应试教育”的附庸产物[21]。

美国 PEM 以 IRT 为指导，基于课程标准构建运动项目测试表现性指标体系，开发丰富多样的“情境性”运动技能测评方案，进而实现对体育运动技能的“真实性”评价，极大地弱化了测评的“应试化”倾向。因此，我国体育中考应广泛纳入“真实性”运动技能测评内容，基于课程标准的目标体系构建统一的运动技能表现性指标体系，据此开发种类丰富的运动技能测评任务或方案，以突破体育中考的“应试化”桎梏，促进学生体育锻炼习惯和终身体育思想的养成，使学生真正掌握一至两项运动技能，提高学生的运动兴趣和锻炼参与热情，最大限度发挥体育中考对学生身心健康发展的长期效益。

3.2 利用测验等值技术，实现体育中考运动技能测评分数的可比性

当前，我国各地区体育中考均含有对学生运动技能模块考核，学生选考一至两项运动技能项目计入总分，主要包括排球垫球、排球发球过网、篮球运球、1 分钟运球投篮、足球运球等内容[22]。暂且不论某一基本技术是否能够代表学生真正掌握这项运动，单从评分标准和计分规则来看，便很难真正体现体育中考的科学性和公平性。如 2020 年 10 月云南省公布的《初中生体育考试专项技能考试内容及分值》中，七年级足球颠球的 0.5 分与篮球 30 秒原地定点双手胸前传球的 0.5 分是否可以等同？相同分数是否意味着难度相同或者说学生需要付出同等时间和精力？随着年级增长，同一项目得分越高是否代表着学生运动能力增强？显然，在没有经过科学论证的情况下以上问题很难给予肯定答案。实际上，在以 CTT 为基础的真分数模型中，受试者能力量表与评价项目难度量表不一致，实测分数并不处于等距量表之上，且由于其对于样本的依赖性很难建立“平行测验”。因此，即使是对同一能力的考核，两个测验分数也难以进行比较。简言之，以真分数模型为基础的体育中考测验中，既不能将不

同运动项目的测验分数进行横向比较，也不能将同一项目的不同测验分数进行纵向比较。简单的分数叠加和对比不仅削弱了体育中考测验的科学性和公平性，同时也无法提供更多大范围反馈和改进信息，由此更进一步加深了体育中考的“终结性”意蕴。

如前所述，测评体系的构建是一项极其复杂而又专业的工作，必须按照规范化、标准化和科学化的操作流程进行。因此，我国体育中考应以省、自治区为单位，在确定测试内容后进行大范围试验和数据收集，利用测验等值技术制定相应的评分标准和细则，实现体育中考分数的可交流性，进一步提升体育中考分数的科学性和公平性。

3.3　研制参数详实的测评工具，提高体育中考分数的精确性和区分度

作为一种升学考试，体育中考的目的不仅在于“以考促练”，提高学生的运动参与和体质健康，还应当兼具考试所具有的竞争、选拔作用。因此，体育中考成绩应有合理区分度且符合正态分布。若大部分学生都能获得高分甚至满分，显然不会引起学生和家长的重视，无法体现体育中考的本质功能，最终极有可能导致体育中考流于形式。然而，调查结果显示，部分地区或学校的体育中考合格率甚至是满分率高达 90%[23]。可见在人人都可拿高分的情况下，体育中考已然沦为“合格性”考试，其效果可想而知。反之，若体育中考分数的差异性和区分度不断提高，其分数必然会引起学生、家长和社会的“锱铢必较”，由此便对体育中考分数的准确性提出更高要求。然而，在真分数模型中所测得的实测分数并不位于等距量表上，同一测试中被测学生必须置于被测对象团体中，根据相对等级和相对位置来评估其能力水平或评分，只有在施测能力水平与测验难度相当的被试者时，才容易获得比较高的测量精度。

以美国 PEM 以 IRT 为基础，对 ATF 研制出的每一个评价工具进行参数估计，将个人能力与项目难度置于同一尺度，最终形成项目难度已知且分布均匀的评价工具库，使得测评者可根据相应难度的测评工具准确定位学生能力，确保测评分数的区分度和精确性。实际上在我国其他学科测评领域，上述技术和方法已经得到运用并取得了突出效果。如大学生英语水平测试，测验者根据难度系数选择试题并形成测验，不仅准确估计受试者真实英语水平，同时也保证每次测验的难度一致。因此，随着体育中考分值的不断上升，精确估计学生体育学习成果和能力便显得尤为重要。我国应充分利用现代教育测量理论的优势，开发参数详实的体育测评工具，确保体育中考测评分数的科学性、严谨性和精确性。

3.4　建立动态体育中考题（项目）库，不断丰富和完善体育中考测试内容

我国体育中考对于正常体育教学秩序的冲击是不言而喻的，这不仅是由于“体能性”“应试化”的测评内容和方式使得体育教学沦为“训练课”，更体现于中考测试内容对体育教学内容选择的束缚。当前我国各省、市公布的体育中考测试内容明显少于《义务教育体育与健康课程标准(2011 年版)》中水平目标所要求和涉及的内容。进一步调查发现，为了“备战”体育中考部分学校只会开设中考体育测试内容所包含的体育课程，不仅限制了学生体育学习内容的可选范围，不利于提高学生体育学习兴趣和动力，而且降低了《义务教育体育与健康课程标准(2011 年版)》对于体育教学的指导意义和价值，很可能导致体育“新课改”多年积累的成果付之一炬。此外，在实施健康中国战略背景下，无论是以《健康中国 2030》政策为代表的宏观设计，还是体育与健康课程标准的中观指引，亦或是学生个人对健康的微观诉求，均体现出新时代国家和人民对健康的重视程度。现阶段，体育与健康课程作为我国健康教育实施的重要平台和载体，体育中考理应纳入对学生健康能力和知识的考核，以此促进学生健康知识的储备和健康生活方式的养成。

实际上，在以 CTT 为基础的测量实践中，因其信效度和误差控制的问题且大多数测试都是孤立开发的，故很难对其进行后续改进和完善。美国 PEM 在 IRT 技术和方法的支持下，遵循题库开发的基本程序和方法，构建内容丰富、科学合理的体育学习成果测评题(项目)库，不仅确保测评项目和内容开发的动态性和可持续性，更实现了对题(项目)库测评工具的不断改进和更新。因此，我国体育中考应遵循题(项目)库开发的基本原理和方法，建立体育中考项目库，不断丰富和完善体育中考测试内容，满足学生对不同运动项目的学习需求，使得“考什么练什么”转变为“练什么考什么”，提高学生体育学习兴趣和动力。同时，也应建立体育中考试题库，采用纸笔测试形式纳入对学生健康知识储备和健康素养的考核，以此促进学生健康生活方式的养成，为深入贯彻落实“新课改”和“健康中国”战略的理念和要求助力。

随着国家和社会对于青少年身心健康问题越来越重视，体育在学校教育中的地位不断提升，各层次、各学段体育考试将成为国家和社会获取体育教学质量有效信息及问责的重要参考指标。鉴于体育中考的制度要求和实践问题，在体育考试“高利害”性越来越

突出的背景下，如何构建科学合理的体育中考测评体系是完善我国体育考试制度的必要前提和必由路径。现阶段我国可借鉴国外优秀经验，充分利用现代测量理论的优势，突破传统体育学习成果测评的现实困境，弥补我国体育学习测量领域的缺陷和不足。在此基础上，还须立足于本土实际，在实践中积极探索体育中考的新方法、新技术、新路径，不断更新和改进体育中考的测试内容、测评技术和计分办法，以建立更加科学、更加合理、更符合现代教育发展趋势的体育学习成果测评体系，为进一步完善我国体育考试制度提供充分经验与技术支持。

参考文献：

[1] 中华人民共和国教育部. 关于印发深化体教融合促进青少年健康发展意见的通知[EB/OL]. (2020-08-31)[2020-11-20]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1779/202009/t20200922_489794.html.

[2] 中华人民共和国教育部. 中共中央办公厅 国务院办公厅印发《关于全面加强和改进新时代学校体育工作的意见》和《关于全面加强和改进新时代学校美育工作的意见》[EB/OL]. (2020-10-15)[2020-11-25]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202010/t20201015_494794.html.

[3] 李小伟，刘亦凡. 中考体育如何在阻力中前行[J]. 人民教育，2020(Z3)：99-101.

[4] ZHU W，RINK J，PLACEK J H，et al. PE Metrics：Background，testing theory，and methods[J]. Measurement in Physical Education and Exercise Science，2011，15(2)：87-99.

[5] 卢荣伟. 项目反应理论在大规模考试试题分析中的应用[J]. 统计与管理，2017，32(10)：50-52.

[6] CHEN W，HAMMOND BENNETT A，HYPNAR A. Examination of motor skill competency in students：Evidence-based physical education curriculum[J]. BMC Public Health，2017，17(1)：222-229.

[7] Society of Health and Physical Educators. PE Metrics：Assessing student performance using the national standards & grade-level outcomes for K-12 physical education[M]. 3rd ed. Champaign：Human Kinetics，2018.

[8] 何毅，董国永. 美国 PEM 体育学习评价体系研究[J]. 首都体育学院学报，2018，30(6)：537-541.

[9] 杜文久. 高等项目反应理论[M]. 北京：科学出版社，2014.

[10] 戴海琦，罗照盛. 项目反应理论原理与当前应用热点概览[J]. 心理学探新，2013，33(5)：392-395.

[11] 郑日昌. 心理与教育测量[M]. 北京：人民教育出版社，2011.

[12] WEIMO Z，CONNIE F，YOUNGSIK P，et al. Development and calibration of an item bank for PE Metrics assessments：Standard 1[J]. Measurement in Physical Education & Exercise Science，2011，15(2)：119-137.

[13] BENEDICT D，JUDITH H P，KIM C G，et al. Development of PE Metrics elementary assessments for national physical education standard 1[J]. Measurement in Physical Education & Exercise Science，2011，15(2)：100-118.

[14] Society of Health and Physical Educators. PE Metrics：Assessing the national standards，standard 1：elementary[M]. Champaign：Human Kinetics，2008.

[15] Society of Health and Physical Educators. PE Metrics：Assessing national standards 1-6 in elementary school[M]. Champaign：Human Kinetics，2010.

[16] CONNIE F，WEIMO Z，YOUNGSIK P，et al. Related critical psychometric issues and their resolutions during development of PE Metrics[J]. Measurement in Physical Education & Exercise Science，2011，15(2)：138-154.

[17] 吴键，袁圣敏. 1985—2014 年全国学生身体机能和身体素质动态分析[J]. 北京体育大学学报，2019，42(6)：23-32.

[18] 杨立远. 体教融合背景下体育中考的历史回顾、现实困境与发展出路——“体育中考”云学术工作坊综述[J]. 体育与科学，2020，41(6)：111-116.

[19] 斯涵涵. 疯狂的应试体育，谁该“吃药”[N]. 健康报，2017-07-13(002).

[20] 徐烨，朱琳. 体育中考的公平诉求及因应之策[J]. 武汉体育学院学报，2013，47(11)：30-35.

[21] 周凰，古雅辉，刘昕. 中考改革背景下学校体育发展的热效应与冷思考[J]. 北京体育大学学报，2017，40(7)：68-75.

[22] 买佳，金光辉，董国永. 利益相关者视角下体育中考执行现状及实施对策[J]. 体育学刊，2020，27(3)：79-84.

[23] 常州中考. 常州市武进体育中考满分率接近 90%. [EB/OL]. (2020-06-07)[2020-11-09]. http://www.wljyyjy.com/ChangZhouZhongKao/364217.html.