心理学探新 2021,Vol.41,No.4,338 -342
PSYCHOLOGICAL EXPLORATION

智能平衡评估系统的概述和启示*

张青华 任 涛 许晓革
(北京教育考试院,北京 100083)

摘 要:智能平衡评估系统是美国目前使用最为广泛的教育评估系统之一。该系统的三大部分——总结性评估系统、中期评估系统和数字图书馆构成一个畅通的教学、反馈、助教与助学的教育循环系统。我国可借鉴智能平衡评估系统的思想和所采用的技术,将相对独立的教育督导、教育研究、考试命题、学校等事业机构在功能上联合起来形成完整的教育链,建构日常教学评估系统,从而提高国内教育的效益,减轻教与学的负担。

关键词:智能平衡评估系统;总结性评估系统;中期评估系统;数字图书馆
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 -5184(2021)04 -0338 -05

1 智能平衡评估产生的背景和特点

智能平衡评估(The Smarter Balanced Assessment)是2010年美国30个州联盟合作创建、开发的一个评估体系和平台(王晓平,齐森,谢小庆,2018)。该评估联盟获美国教育部力争上游(Race to the Top)基金1.76亿美元拨款用于评估系统的建立和开发。智能平衡评估体系的创建源于教育者们意识到当时大多数的教育评估体系是孤立、过时的,不能提供一套有效的工具来帮助教师和学生提高教与学。他们希望创建出一个最好的评估体系,采用一种有效、公平、可靠的技术与方法对学生进行评估,并将有意义的评估结果提供给教育者、学生和家长,从而培养出更多能为未来的大学学习或就业做好更充分准备的学生。

智能平衡评估系统包括总结性评估(Summative Assessment:The End - of - Year Test)、中期评估(Interim Assessments)和数字图书馆(Digital Library)三部分。由4700多名州教育机构人员、教师、高校教授和其他教育工作者参与创建和开发,历时5年(2010年开始研发,至2014年夏季、2014年冬季、2015年春季数字图书馆、中期评估系统和总结性评估系统分别开始投入正式使用)初步完成(Smarter Balanced,2019a)。教师参与系统开发的每一个阶段,如编写和审核测试题、数字图书馆的创建和核查工具的开发、制定成就水平用于确定每名学生达到何种学业水平标准等。智能平衡评估系统具有的特点:在线实施,对学生进行可定制化测试;测试项目主要考查批判性思维、写作和问题解决等重要的技能;用于评估的每道试题都经过偏差和敏感性审核,确保对测试所有学生都公平;帮助教师专业化发展,并提供课内评估工具等。目前该套系统已成为美国历史上使用最为广泛的教育评估系统之一。

2 智能平衡评估系统的结构

智能平衡评估系统主要由三部分构成:总结性评估系统、中期评估系统、数字图书馆。总结性评估是在年终评估3~8年级和11年级学生英语和数学学科的学业成就与成长;中期评估用于支持全学年的教与学;数字图书馆是一套工具和资源,可用于以课堂为基础的形成性评估过程。这三个部分组合成一个辅助教师教学、学生学习、对教与学评估反馈的有机整体。

2.1 总结性评估:年终测试(Smarter Balanced,2019b)

总结性评估是在学年末对学生英语和数学的学业成就和进步情况所进行的评估。评估的对象是3至8年级和11年级的学生。评估的内容有英语(主要包括阅读、写作、听力、研究)和数学(包括概念和程序、问题的解决、模型和数据分析、推理)。评估的方式有两类:计算机自适应测验和表现性行为任务(它在计算机上完成,但不是计算机自适应测验形式,见图1)。表现性行为任务测量学生的批判性思维和问题解决的能力,它要求学生应用他们的知识和技能解决复杂和实际的问题。它们可以很好地将单个主题或场景与问题和活动(活动旨在测量能

* **基金项目**:北京市教育科学"十三五"规划重大课题(2019年度)"中高考改革背景下北京市教学、考试、招生一体化研究"(CAMA19055)。
通讯作者:张青华,E - mail:zhangqh@bjeea.cn。

力,如深度理解、写作和研究技能,以及复杂分析等)紧密结合。而传统评估问题无法充分评估这些能力。评估的结果一方面能准确地反馈学生的学业成就(学年结束时学生掌握了多少知识和技能)和学生的成长(学生比上一个学年进步了多少),另一方面可向教育行政部门对学校、地区和州的教育问责提供相关的信息。目前有超过220所学院和大学采用智能平衡的高中总结性评估结果作为大学入学和学分课程的依据。

艺术日:你正在帮助四年级的学生组织一个艺术节。有三项活动:绘画、陶器、粉笔艺术。你有两项任务需要完成——帮助创建供应列表和艺术日活动时间计划安排。
任务1:供应清单。
您需要确保每项活动有足够的所需要用品供大家使用。您将使用以下信息为您的班级创建用品列表。您所在的班级有24名学生,每名学生需要:绘画活动2把涂料刷、制陶器活动3磅粘土、粉笔艺术活动5支粉笔。
任务2:活动时间安排规划
您需要依据以下提供的信息制定艺术日的时间安排
1、艺术日的活动从上午9点开始,下午2点结束;
2、您的班级全体成员都将轮转参加三个活动;
3、活动休息时间至少10分钟;
4、活动休息时间和午餐时间共1小时;
5、三项活动(绘画、制陶器、粉笔艺术)的时间不必相同,但每项活动至少30分钟或更长。

图1　四年级数学表现性行为任务示例

总结性评估采用了计算机自适应测验(Computer Adaptive Testing,简称CAT)教育测量技术,CAT可根据每名学生的作答调整试题的难度,即如果学生作答正确,则下一道试题会更难,如果学生作答错误,下一道试题会更容易,为每一名学生提供适合其水平的测试试题,因而可以快速测试出学生所掌握的知识和能力水平,从而实现对每一名学生的知识和能力水平进行精准的测试。

2.2　中期评估(Smarter Balanced,2019c)

中期评估是教师可全年用于检查学生学业是否进步,并通过反馈信息改进教学、帮助学生达到大学和职业准备标准的评估工具。中期评估具有的特点:试题质量高,测试内容涵盖了共同核心国家标准描述的深度知识范围;可测试3~8和11年级英语和数学学科的知识和能力;采用测验内容固定的在线测试形式(但不是计算机自适应测验形式),与总结性评估具有相同量尺;教师可获得测验试题以及学生在测验试题上的作答反应,这使教师在教学过程中能了解学生学习的相对强项和需要提升的方面,更好地为每名学生提供个性化的指导,满足每名学生个体需要;测试灵活,可满足教师日常教学需要等。中期评估有两部分构成:中期综合评估(Interim Comprehensive Assessments,简称ICAs)和中期模块评估(Interim Assessment Blocks,简称IABs)。

中期综合评估主要用于确定学生在学年内经过一段时期的学习后所掌握的英语、数学学科的知识和技能。通过中期综合评估可向学生提供英语和数学学科的总体表现情况。根据学生的学业评估结果将学生分为四类:不满足标准、接近标准、达到标准、超过标准。评估采用计算机化测试,测试的时长需3至4小时,可采用标准化或非标准化方式进行测试,测试的内容和测试的标准与总结性评估相同。

中期模块评估主要用于测试学生数学和英语学科的有关概念集,为教学提供更为详细的反馈信息。它关注于对特定内容的测试,例如:测量和数据、分数、阅读信息文本等。不同的测试年级其测试的试题数量和内容范围也不同,试题数量为4至18道,测试的每一个内容领域都包括一项表现性行为任务。评估的结果将学生分为三类:高于标准、接近标准、低于标准,同时可向教师、家长和学生提供有关信息,如学生已经掌握了哪些概念,有哪些方面还需要额外的帮助。教师可在全学年的教学过程中根据课程教学进度灵活使用中期模块评估。

2.3　数字图书馆(Smarter Balanced,2019d)

数字图书馆是教育者为教育工作者提供的一个在线收集教学和专业学习的资源库。这些资源符合美国教育的共同核心国家标准,并帮助教师将这些资源应用于日常教学的形成性评估过程中,以改进教与学。数字图书馆的主要功能有:为不同的学习者提供差异化教学指导;提高教育者对评估的理解;帮助学生自学;设计专业的发展机会;为专业学习机构提供材料。为了实现数字图书馆的功能,课堂教育者与智能平衡评估服务商合作,审核中期模块评估,确定每一种成就水平相对应的知识和技能,并且确定相对应的数字图书馆资源。具体过程为:(1)教育者按学科和年级分类审核学生中期模块评估资

料,以确认每个成就水平学生能正确回答的问题;(2)审核后,教育者对每种层次水平学生可能知道的或能够做到的内容和技能进行描述;(3)教育者从数字图书馆搜索资源,为每种成就水平的学生提供通向成功的指导。数字图书馆资源包括:课堂活动、任务、作业、课程计划,每种特定水平学生都可获得这些资源,并能获取对下一个等级水平资源的了解;(4)根据他们的搜索,教育者为每种成就水平学生推荐一组资源。

数字图书馆这些功能的实现得益于教育专家的通力合作,将智能平衡中期模块评估的学生表现与数字图书馆资源相链接。因为数字图书馆将中期评估结果与其它课堂评估和专业意见连在一起,教育者可以通过使用这些链接找到与学生的需要相一致的相关有用的教学指导,从而协助教师促进学生的学习和成长。由于数字图书馆提供了部分教育者推荐的数字图书馆资源,这些资源可以补充课程和其他课堂活动,但它们并不意味着可以取代课程或教学计划。许多资源可以直接被使用,另外一些可能需要调整以适用于特定的课堂和个体学生。所以在日常教学当中教师可以根据实际教学情况决定如何使用数字图书馆资源来帮助他们的教学。

3 智能平衡评估系统的报告分数体系(Smarter Balanced,2019b)

智能平衡评估系统在总结性评估和中期评估中采用(1,0)计分项目、表现性行为任务对学生进行测试,使用双参数逻辑斯蒂克模型(two - parameter logistic model)和广义分部评分模型(generalized partial credit model)对测试项目难度和学生的能力进行估计,量尺以学生能力的均值为0,标准差为1作为单位,并将学生能力值转换为四位数的量尺分(量尺分 = 能力 × 斜率 + 截距)。智能平衡评估学生的数学量尺分 = 学生能力 ×79.3 +2514.9、英语的量尺分 = 学生能力 ×85.8 +2508.2。并采用共同锚题的等值方法建构跨年级的垂直量尺和题库。

智能平衡评估系统对学生进行评估后报告的结果主要有两类指标:量尺分(scaled scores)和成就水平(achievement levels)。量尺分是学生的评估总分,它是一个分值范围为2000至3000的连续量尺分,且随着年级而增加。量尺分可用于解释学生目前的成就水平和他们不同时期的成长,同时也可用于描述学校层面和地区层面的教育行为变化,以及不同学生群体之间的成就差距。中期模块评估、中期综合评估和总结性评估的量尺分都在同一垂直量尺上(即所有年级学生行为都在相同量尺上进行报告和解释)。

智能平衡评估系统的学生成就水平是根据学生的量尺分把学生分为三种成就水平(中期模块评估把学生的成就水平分为低于标准、接近标准、高于标准)和四种成就水平(中期综合评估和总结性评估把学生的成就水平分为:水平1、水平2、水平3、水平4,或者称为不达标准、接近标准、达到标准、高于标准)。成就水平的描述是由教师和大学教师来编写。每一种水平都描述了学生在知识和技能上有哪些具体的表现。例如,11年级在"水平3"或以上的水平意味着学生已经准备好接受入门级的、可转换的、有学分的大学课程。

量尺分与成就水平之间的对应关系见表1示例。确定每种成就水平对应量尺分的分数阈值是一项重要的工作,智能平衡评估系统采用书签法(BOOKMARK方法)对每类成就水平进行分数阈值的确定。该方法首先是成千的K-12教师、高校教师、家长和其它相关方参与成就水平界定的建议;其次,由教师和其它利益方组成的按年级分类的工作组,对每个年级的各种成就水平的分数阈值进行讨论并提出建议;再次,跨年级审核委员会审查所有年级的建议,并考虑合理的划界分数体系。在确定两者关系的同时,与英语学习者和残疾学生一起工作的教育工作者也参与其中,以帮助确保成就水平对所有学生都是公平和适当的。

表1 中期综合评估的数学成就水平与量尺分对应关系表

年级	水平1	水平2	水平3	水平4
3	<2381	2381 - 2435	2436 - 2500	>2500
4	<2411	2411 - 2484	2485 - 2548	>2548
5	<2455	2455 - 2527	2528 - 2578	>2578
6	<2473	2473 - 2551	2552 - 2609	>2609
7	<2484	2484 - 2566	2567 - 2634	>2634
8	<2504	2504 - 2585	2586 - 2652	>2652
11	<2543	2543 - 2627	2628 - 2717	>2717

(注:参考Smarter Balanced,2019e)

智能平衡评估的结果不仅可用于说明学生目前成就水平,而且还可用于帮助学生、教师和家长评估学生的学业进步。但在使用评估结果时需要注意,不应简单使用成就水平来描述一个学生的学业成就,还需要结合量尺分、成长模型,以及学生作品集来共同评估学生、学校的学业进步。

4 智能平衡评估系统的应用(Smarter Balanced,2019e)

4.1 智能平衡评估系统对于学生的应用

智能平衡评估系统对学生有三种评估类型:中期模块评估、中期综合评估、总结性评估。每类评估都会给学生提供相应的评估报告,学生通过评估及

评估反馈(评估报告),使其了解自身在评估的学业内容上哪些掌握了,哪些方面需要加强,自身在不同年级阶段的成长。学生的中期模块评估报告内容包括:(1)学生的信息:姓名、年级、学校、学生所在的地区和州;(2)报告的名称(例如:“2017年8年级数学中期模块评估”);(3)评估内容名称(例如:8年级数学方程中期模块评估);(4)学生的量尺分和测量误差范围;(5)评估的日期和学生的成就水平类别;(6)常见的问题;(7)有关中期评估的相关信息和其他资源。

智能平衡评估系统学生中期综合评估报告单内容包括:(1)报告单的名称、科目、年度(例如:中期综合评估报告单、数学、2017~2018年度);(2)学生信息:姓名、年级、学生所在的地区和州;(3)有关学生成就的信息:量尺分、量尺分的测量误差范围、成就水平、以及其它可能的成就水平(每种成就水平对应的最大、最小量尺分);(4)学生成就水平的描述(例如:“某某学生取得了学业的进步,并达到了8年级的数学标准,可以进入下一阶段的学业学习”);(5)学生每一项被评估内容的成就水平说明(例如:某同学具有一定的应用问题解决技能和策略,并用学科知识解决数学特定问题的能力);(6)其它信息。

智能平衡评估系统的总结性评估采用计算机自适应测验技术,根据每一名学生的学业水平提供相应难度的测试,更高效、更精准地测量出学生的水平。测试向学生提供学科的总分和各分项内容的分项分数,并对总分和各分项分数的内涵进行解释(如学生达到何种成就水平、掌握了什么知识和具有何种能力等)。

4.2　智能平衡评估系统对于教师和教育工作者的应用

智能平衡评估系统为教师和教育工作者提供的应用是多层次、多方位的:一方面,10个州的250多所学院和大学把评估结果作为决定学生是否准备好接受学分课程的一个因素;另一方面,3~8和11年级的教师可以通过智能平衡评估系统的总结性评估、中期评估和数字图书馆对学生年末、日常教学等进行学术检查,并获取有关针对各类学生的教学指导信息,从而帮助学生取得成功。以下用一个中期评估的示例来阐述智能平衡评估系统在教师中的应用。

加西亚老师教授3年级某班的英语,该班级有20名学生。以一次英语阅读的中期模块评估为例,通过评估报告系统加西亚老师可以了解到三个层面的信息:第一学生群体层面信息。本次班级学生的英语阅读平均分为2380,误差范围为±29。有5%的学生高于标准、65%的学生接近标准、30%的学生低于标准。由于每一个中期模块评估的报告系统都与数字图书馆列表连接,列表可以提供每一个模块对应的学生报告类别(高于标准、接近标准、低于标准)相应的教学指导资源(例如,针对高于标准的学生提供更丰富和扩展他们技能的教学指导、根据学生的需求提供有差异性的教学指导等)。第二群体在项目层面的信息。评估报告系统可以提供中期模块评估中每道试题的测试要求、目标、难度和有关的评估标准,以及获得满分学生的比例和各个分点的学生比例。这使老师能够快速识别学生在哪些评估试题上表现好,哪些试题上表现不足。第三学生个体层面信息。一方面,通过评估报告系统,加西亚老师能够根据每种报告类别的检索获得每名学生的情况,并对不同类别学生做出差异化的教学指导。例如,对于评估结果高于标准的学生,她可以再结合学生的课堂表现、家庭作业和其它的观察判断学生是否已经掌握了课堂所教的知识和技能,如果掌握,则对这类学生不再提供此评估内容的教学指导。从而她可将注意力集中在低于标准的学生,为这类学生提供有针对性的额外教学指导。通过中期模块评估,可以帮助教师调整对不同类别学生的教学指导,从而改进教师的教学与学生的学习。另一方面,教师可以查看每一名学生在评估试题上的作答信息(如多项选择题,可以获得每名学生的选项信息),如果有几名学生选择了相同的错误选项,教师可以识别学生的反应模式——揭示出学生们存在的共同错误点和理解偏差,从而改进课堂的相应教学。

5　启示与思考

2018年9月10日我国召开的全国教育大会提出深化教育领域综合改革,着力形成充满活力、富有效率、更加开放、有利于高质量发展的教育体制机制。教学评估是日常教育教学中的重要一环,科学、高效的评估能够提供及时的教学反馈,有利于促进教师教学的改进和学生的学习,实现科学的减负增效。

目前我国有各种层面的学业评估,如国家级的教育质量监测、地方性的期中和期末考试。同时我国的教育体系中设有学校、教育考试命题、教育督导、教育研究等事业机构,然而这些事业机构由于在体制上的独立,造成教学、考试与评估、督导、教学研究各自为政,相互独立的现状。本文介绍的智能平衡评估系统通过可选择性的和灵活的中期评估对学生日常学习进行评估与反馈,通过总结性评估在学年末对学生整个学年的学习结果进行评估,通过数字图书馆使教师能根据学生中期评估和总结性评估的反馈结果获取相应的资源改进和调整自身的教

学,从而帮助和促进学生的学习。中期评估系统、总结性评估系统和数字图书馆三者构成了一个畅通的良性循环系统。结合我国现实的国情,我们能否将在体制上相互独立的教、评、研、督教育机构进行联合,形成完整的教育循环体系链,从而提高教育的质量与效率,科学地减轻教师和学生的负担。

2014 年我国各省、直辖市分批进入新一轮的高考改革。新的高考改革带来新的高中教育教学方式的改变——分层走班制。在这种新形式下,对不同层次班级教学进行科学、有效评估是日常教学面临的新问题。智能平衡评估系统采用在线测试方式(包括计算机测试和计算机自适应测试方式两种),不仅可以精准测量出不同层次学生的真实水平,而且评估结果的反馈更省时、便捷。计算机自适应测验技术目前已发展比较成熟,能否建立采用自适应测验技术的国家级、省级(或跨省联合)的以新课程标准为依据的学业评估系统,用于日常教育教学评估,为日常一线教师的教学、学生的学习提供及时的反馈,实时有效地帮助教师改进教学、促进学生学业成长。

此外,目前我国对学生学业的评估主要集中于"是什么"——评估学生当前知道什么和能做什么,而很少对学生个体学业"成长"的变化进行评估。智能平衡评估系统对于 3 ~ 8 和 11 年级学生的英语、数学学科的学业评估分别建立在可比较的同一个量尺上,这样使每名学生个体能够了解自身学业的成长与发展。我们的教育评估与督导部门能否建立起一套评估体系,纵向地评估和监测学生个体的学业发展,这不仅对于学生而言,使其能了解自身学业成长,同时也能够评估和监测国家、地区的教育质量发展状况,有利于国家教育的长远发展。

鸣谢:在本文形成过程中得到北京语言大学教育测量研究所谢小庆教授的大力支持与帮助,特此鸣谢。

参考文献

王晓平,齐森,谢小庆.(2018).政府怎样支持教育科学研究?——美国联邦政府的一些做法.*教学管理与教育研究*,*19*,121 - 123.

Smarter Balanced. (2019a). *What is smarter balanced*. Retrieved March 10,2019, from https://www.smarterbalanced.org/about/

Smarter Balanced. (2019b). 2016 - 17 - *summative - assessment - technical - report*. Retrieved March 10,2019, from https://www.smarterbalanced.org/assessments/development/

Smarter Balanced. (2019c). *Interimassessments overview*. Retrieved March 10,2019, from https://www.smarterbalanced.org/assessments/interim - assessments/

Smarter Balanced. (2019d). *Digital - library - connections - overview*. Retrieved March 10,2019, from https://www.smarterbalanced.org/educators/the - digital - library/

Smarter Balanced. (2019e). *Interim assessments interpretive guide*. Retrieved March 10,2019, fromhttps://www.smarterbalanced.org/assessments/interim - assessments

Overview and Enlightenment of the Smarter Balanced Assessment System

Zhang Qinghua Ren Tao Xu Xiaoge

(Beijing Education Examination Authority, Beijing 100083)

Abstract: The Smarter Balanced Assessment System is one of the most widely used educational assessment systems in the United States. The three major parts of the system - summative assessment system, interim assessments system and digital library constitute a smooth educational cycle system of teaching, feedback, teaching assistants and learning assistants. Through learning from the idea and technology of the Smarter Balanced Assessment System, we can combine relatively independent institutions such as educational supervision, educational research, examination proposition, schools and so on to form a complete educational chain in terms of functions, and construct an educational assessment systems used on daily teaching, so as to improve the efficiency of domestic education and reduce the burden of teaching and learning.

Key words: the Smarter Balanced Assessment System; Summative Assessment; Interim Assessments; Digital Library