北京大学
中国教育财政科学研究所

简　　报

怀仁怀朴　唯真唯实

2013 年第 1 期（总第 75 期）　　2013 年 1 月 15 日

开发新的评价工具：对幼儿园过程性质量进行表现性评价

黄晓婷　宋映泉*

[**摘要**]科学的学前教育质量评价工具可以为促进幼儿园质量提高提供重要信息，具有重要的实践意义和理论价值。本研究运用威尔逊四步法构建了《幼儿园教学观察表》，尝试对幼儿园教师课堂互动中的态度和行为进行表现性评价。观察员对某省来自 120 所幼儿园近千名幼儿教师进行了观察评分。初步结果表明该观察表具有一定的信度和效度，可以以此为基础进一步发展幼儿园过程性质量的表现性评价工具。此外，本文详细介绍了观察表的制定过程和信效度验证，旨在促进我国学前教育质量评价工具的科学构建、完善评价标准的进一步讨论和研究。

[**关键词**]幼儿园过程性质量；表现性评价；威尔逊四步法；信效度

* 作者简介：黄晓婷，北京大学中国教育财政研究所助理研究员；宋映泉，北京大学中国教育财政研究所副研究员。

一、研究背景

自 20 世纪 80 年代起，学前教育在世界各国日益受到重视。国际上的多项研究表明，学前教育对儿童的语言、智力、情感等发展，乃至对打破贫困的代际传递、降低犯罪率、促进社会公平都有着重要而长远的影响。优质的学前教育能够培养儿童各方面的能力，使他们在未来取得更好的学业成绩，获得更高的经济收入和社会地位，促进社会的发展。反之，质量堪忧的学前教育不但不能带来这些益处，而且还会对儿童的身心发展造成长期的负面影响。[1] 因此，如何提高学前教育的质量已成为这一领域的研究热点。

近年来，我国的学前教育规模也急速扩大。据《中国教育统计年鉴》的数据显示，我国全职幼儿教师的人数从 2001 到 2010 年几乎翻了一番。2010 年的《国家中长期教育改革和发展规划纲要》更提出了到 2020 年基本普及学前教育的目标，进一步加大了我国学前教育的发展力度。与此同时，学者们也越来越清晰地认识到在普及学前教育的过程中，我们还必须兼顾质量的保证和提高。[2]

怎样定义学前教育的质量一直以来都是研究者们争论的话题。各个国家和地区都根据自己的历史、经济、文化、政策等社会现状，对其进行诠释。但万变不离其宗，国际经合组织（OECD）的研究发现，各种对于学前教育质量的定义都包涵“结构性质量（structural quality）”和“过程性质量（process quality）”两个部分，并以儿童的各项发展为最终目的。[3] 结构性质量包括硬件设施、资源、教师资质、师生比等因素。这些因素相对稳定，往往可以直接观察到。而过程性质量则关注教师与儿童之间互动的本质，也包括教师与家长、儿童之间的互动，以及教师教学技巧、幼儿园的领导和管理等。这些因素与儿童每天在幼儿园里的体验息息相关，但它们的浮动变化较大，评估起来比较复杂。结构性质量和过程性质量这两方面紧密相关，对儿童的成长发展都有着至关重要的影响。要提高学前教育质量，就必须考虑到多个重要因素。[4]

我国关于学前教育质量的研究已经有了一些积累和经验。很多省、市教育部门都先后颁布过幼儿教育的评价标准，这些标准更多地关注结构性质量的因素。[5]对过程性质量的研究，如“幼儿园一日活动质量评价”等[6-8]，往往注重于对教师的行为和教学活动的评价。而聚焦师生互动，体现以幼儿为本教育理念的评

价工具及其研究却寥寥无几。这一方面是因为我国关于幼儿教师的角色、职责、价值和标准的研究相对滞后，另一方面是由于互动等因素的评价难度大，需要通过表现性评价（performance assessment）来实现。表现性评价即通过实地观察被测人的日常行为表现来进行评价。这类评价工具的开发和实施需要投入大量的时间和人力成本，而且在我国少有经验可循。

本研究瞄准过程性质量中的互动因素，运用威尔逊四步法构建了《幼儿园教学观察表》（下简称“观察表”），尝试对教师在课堂互动中的态度和行为进行表现性评价，并对这一评价工具作了初步的信、效度验证。本文的目的并非总结性地评价某地或某些幼儿园的互动因素，而是期待抛砖引玉，为未来更多对过程性质量的评价和研究提供一些思路和经验。

二、评价工具的构建

威尔逊四步法是由美国教育学家Mark Wilson从自己多年的研究经验中提取出来的一套设计构建评价工具的方法，包括“构建图—》问题设计—》表现汇总—》建模分析”四个模块。这种方法目前已受到国际教育测量界的广泛认可。[9]四步法的具体步骤如下：

（一）构建图

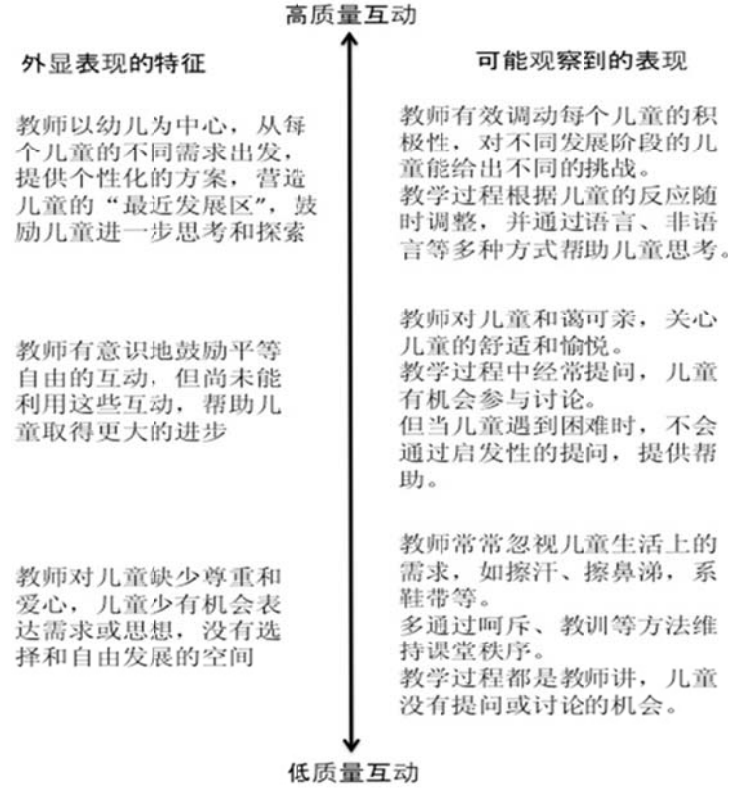
第一步，“构建图”，即根据现有的理论和实证研究结果，勾勒出一幅构建图（construct map），如图 1 所示。画构建图首先要明确评价的内容，即“构建（construct）”的定义是什么，包含哪些内容。然后界定不同发展水平的人或事业分别有哪些外显的行为或表现。最后思考什么样的问题可以反映出这些不同的发展水平。

本研究中的构建即幼儿园中互动因素的质量，通常指教师是否能利用语言或非语言的交流，为儿童智力、情感等方面的发展提供最有效的支持和帮助。[10]依据理论和国际上多个较为成熟的过程性质量评价量表，如《幼儿早教课题观察量表》（*ECCOM，Early Childhood Classroom Observation Measure*）[11]、《课题评价评分系统》（*CLASS，Classroom Assessment Scoring System* ）等 [12]，观察表计划包含“互动环境”、“语言互动”和“非语言互动” 三方面内容。互动环境，

即利于儿童学习、活动和交流的物理及心理环境。语言互动，指教师与儿童间、儿童相互间的语言交流。非语言互动，顾名思义，即教师与儿童间、儿童相互间通过表情、肢体等方式的交流。

图 1 为本研究的构建图。双箭头的直线表示互动质量，越靠上方表示互动质量越好。图中直线的左侧描述不同水平的外显表现的特征，右侧列举在评价过程中可能观察到的表现。本文将互动的水平精简为高、中、低三个层次。较高质量的互动，以幼儿为中心，从每个儿童的不同需求出发，提供个性化的方案，营造儿童的“最近发展区（zone of proximal development）[13]”，让儿童得到最大的发展空间。中等质量则表现为有较多平等自由的互动，但教师尚未能利用这些互动，帮助儿童进一步思考探索。而质量差的互动，则完全以教师为中心，极度缺乏沟通，或居高临下，儿童没有自由发展的空间。右侧列举的表现包括观察教师在课堂秩序管理、讲解和讨论中对儿童的态度和行为。虽然过程性质量中“互动”还包括教师与家长、教师与园长等，但由于实际操作中受到的时间、资源等的局限，本研究只将评价范围局限在课堂教学过程中能够观察到的教师与儿童及儿童相互之间的互动上。

图 1 幼儿园互动因素构建图



（二）问题设计

有了构建图之后，第二步“问题设计”就可以依据蓝图，确定评价中将使用的具体问题。本研究的一个重要挑战是必须在一个课时内（约 25 分钟）完成观察和评价。考虑到这个制约条件，《幼儿园教学观察表》最终只保留了 11 个问题。其中有 4 个问题针对“互动环境”，如教室空间安排是否方便儿童取得需要的物品、教材等；教师对待儿童的总体态度是否亲切、循循善诱；课堂秩序维护的方式和结果等。有 4 个问题针对“语言互动”，如教学过程中的师生问答；儿童间的讨论交流等。最后 3 个问题针对“非语言互动”，如教师对儿童需求的体察，教师使用肢体语言鼓励或安抚儿童等。具体的问题因尚需保密，本文暂不一一展示。

（三）表现汇总

第三步，“表现汇总”即通过专家讨论或试测等方式，确定评分标准，保证各种可能出现的答案都可以有对应的、合理的得分，避免在正式的评价过程中出现“这种情况既可以得 3 分又可以得 4 分”的模棱两可的状况，或“这种情况得几分都不太合适”的无所适从的情况。

本研究观察表中的问题最初设计均有 4 个选项，对应得分为 0 到 3 分，得分越高表示互动质量越好。如观察教师对儿童总体态度一题，对儿童冷淡、严厉，经常忽略对儿童的问题或要求，或采取更负面的态度（如粗暴对待）得 0 分；态度温和，但严肃刻板，以发号施令为主的情况得 1 分；态度亲切，大部分时候保持微笑，能通过语言或轻抚等动作表现爱心的情况得 2 分；态度热情，和儿童保持平等交流，尊重和理解儿童的多种需求，鼓励儿童主动发问等表现得 3 分。观察员依据“主旋律”表现进行评分。在试测修改的过程中，研究者们发现有些问题的情境在短短一课时中没有出现，因此对其中 4 个问题增加了“未能观察到”的选项。这种情况后期作为缺失数据处理。

（四）建模分析

最后一步，“建模分析”指在实施评价后，运用项目反应理论（IRT, Item Response Theory）的模型分析结果，同时验证量表的信度和效度。在本研究中，

我们运用分部计分模型（Partial Credit Model，PCM）[14]分析结果。PCM 适用于每题有多个分值，难度随分值的增加而增加的情况。国际上的很多测试、问卷（如托福，GRE，PISA）等都使用这一模型进行分析。较之经典测量理论（Classical Test Theory）或其他传统的分析方法，PCM 有着无可比拟的优势：它对受试者的能力估值不受问题特性影响，不会因为问题难能力估值就低，问题容易估值变高。同样，它对问题难度的估算也不因为受试人群的特性而改变。因此，评价的结果可以再次映射到构建图上，为建立或完善评价标准提供有价值的信息。另外，它不以总分为依据，而是针对每一题、每一个得分项与受试者之间关系进行分析。它对精度的估算也会因人而异、因题而异，比传统方法中的一概而论更符合现实情况。

运用 IRT 专业软件 ConQuest[15]，研究者可以同时得到信度系数、数据与模型适配指数、每一题的难度、每一个得分项的难度，及单题得分与总分相关性系数等。这些信息可以用于验证量表的信度和效度。信度越高，评价结果的误差越小；效度越好，则量表中的问题越有的放矢，正中靶心。倘若信度或效度不能符合要求，那就需要回到第一步，完善理论基础，修改构建图，重新再完成四步过程。

三、评价的实施

表现性评价不同于传统的纸笔考试或问卷调查，需要由评分员仔细观察被测人在具体场景中的表现，然后中立、客观地作出评分。因此，在进行评价前，所有观察员都接受了培训。先由量表设计者逐条解释问题和得分项，强调观察和评分时的注意事项，如观察时要做“隐形人”，不影响课题内的任何活动；判分需在观察结束后立即完成等。然后大家一起观看幼儿园的课堂教学录像。观察员对录像中的互动质量独立进行评分。完成评分后，再和专家评分对比，讨论分歧，力求达成共识。最后再进行两次评分练习，以期提高评分的准确性。

实测的样本为随机抽样的某市 818 名幼儿园教师，其中女教师 768 名，占绝大多数（94.2%）。农村户口教师 295 名，占 36.1%。超过半数的教师为高中或中专毕业（53.1%），仅有 10.9%的教师有本科或以上文凭。这些教师的平均工作经验为 7.3 年，无任何职称的占 56.7%。本研究进行时，在托班任教的教师有 5 名，

小班 210 名，中班 214 名，大班 213 名，学前班 117 名。

这些教师被随机分为两组，实验组的教师连续两个周末参加了提高教学质量的培训，培训时间共 32 小时。培训由专业培训机构承担，内容以联合国儿基会的《爱生幼儿园教育指南》为基础，并突出教学中的互动技巧。548 名教师在干预实验后两周再次参与了评价，其中 375 名为实验组，173 名属控制组。样本损耗率较高的原因是观察员在实验后第二次实施评价时，由于课程安排、个人原因等，无法在规定日期内对这些教师进行观察评价。因此，本研究培训前后共观察了 1366 课时。总数据量完全可以得到稳定的统计分析结果。

四、初步结果

根据分部记分模型（PCM），题目难度是分布为在-4 到 4 之间的一个连续变量。数值越小表示水平越低。分析结果表明数据与 PCM 模型适配良好。两次的评价对教师互动质量的估算都呈正态分布。表 1 总结了干预前后两次评价的总体情况：干预后平均水平有小幅度提升，在统计上有显著意义。干预后，水平分布的标准差略有缩小，从最低到最高的跨度缩小了近一半。

表 1《幼儿园课题观察表》两次评分结果

	样本量	平均分	标准差	最低分	最高分
干预前	818	0.01	0.56	-3.43	3.18
干预后	548	0.25	0.53	-1.63	2.22

对干预前后都参加了评价的 548 名教师，表 2 展示了控制组和实验组在干预前后的评价结果。实验组提升了约 1/2 个方差，而控制组则仅提高了近 1/6 个方差，两组差异在统计上也具有显著意义，效应值（effect size)也较大。因此，短期培训取得了一定的效果，但长期效果还有待证实。

表 2 控制组和实验组在干预前后的评价结果

	样本量	前测均值	后测均值	平均增涨	增长的标准误
控制组	173	0.06	0.17	0.11	0.05
实验组	375	-0.01	0.29	0.30	0.03

此外，结果显示对教师来说难度最低的题为“在活动中为保证儿童安全，采取防范措施”，而最难的题则是“通过语言交流鼓励幼儿表达复杂的思想或回答较难的问题”。总体的平均水平接近 0，对应构建图，抽样调查到的课题互动质量为中等。各题中难度接近 0 的得分项可以为研究者大致描述平均水平教师的表现，而难度超过 0.5（约一个方差）的则粗略勾勒高水平教师的表现。这些信息有助于修改和完善未来的培训内容、培训目标，还可以为建立更完整成熟的过程性质量评价体系提供实证依据。

五、评价工具的信、效度

运用 PCM 模型，计算观察表的信度系数为 0.66。在表现性评价中，这样的信度尚在可以接受的范围，但还有很大的提升空间。例如，在其他条件不变的情况下，可以通过增加观察表的问题数量来提高信度。如果增加至 22 题，根据 Spearman-Brown 公式，信度可达 0.8。对观察员加强培训也可以减小误差，提高信度。另外，修改问题和评分标准，使其更容易“对号入座”；增加观察时间，让观察员能对互动表现有更准的把握等举措，也能在一定程度上提高信度。

对观察表效度的验证，本研究主要从内容、内部结构和外部关联三个方面进行。首先，在内容方面，观察表的维度和问题在国际知名的同类表现性评价量表中有着不同形式的体现，同时切合联合国儿基会的教育指南，评分的标准亦符合儿童发展理论和国际上主流的教育理念。[11-13]

结构上，每个问题的得分与总分的关联都大于 0.23，可见所有问题都指向正确。每题中，与从 0 至 3 的得分项对应的 3 个步骤难度（step difficulty)[14]依次递增。这说明每题评分标准的设定也正确合理。由于“互动环境”、“语言互动”和“非语言互动”这 3 个维度的题量都较少，本文没有进行多维度分析。如果题量充裕，内部结构合理，不同维度间应该有较高的相关性。

最后，研究者通过比较控制组和实验组的变化，验证了该量表的外部效度。由于培训的内容突出过程性质量，理论上接受培训的实验组应该比控制组取得更高的得分。分析结果表明，两组教师在观察表中的得分符合预期，培训有一定效果，水平的提高幅度也符合短期培训的近效预期。此外，控制组得分也有所提高，最可能的原因是虽然培训人员要求实验组教师保密，但本研究未签署保密协议或

采取其他保密措施，仍可能由于同事间的交流导致溢出效应。

六、总结和讨论

本文的研究结果表明，通过威尔逊四步法构建的观察表具有一定的信度和良好的效度。该量表初步评价了某市幼儿园过程性质量中的互动因素，并对实验中的短期培训效果进行了初步评估。在此基础上，研究者还可以进一步完善量表，制定能更全面反映学前教育过程性质量的表现性评价量表，并为修改和完善幼儿园质量评价标准提供有效信息。

在过去的研究中，对过程性质量的探讨并不多见。表现性评价的复杂性常常使研究者和实践者望而却步。本研究表明，只要依据理论，通过认真、仔细的设计和实施，我们完全可以对过程性质量作出较为准确的评价。本文的主旨，就在于抛砖引玉，希望能引起更多研究者和实践者在这方面的讨论。

由于受时间和资源等因素的限制，观察表题量较少，观察时间很短，因此对课堂内师生互动的质量只能作出粗略的描绘和评估，信、效度都有待进一步提高。在后续研究中，我们还将对教师在实施不同的课程内容时、或不同情境下对儿童的鼓励、安抚、启发等表现，进行更细致地观察和评价。此外，过程性质量还包含教师之间、教师与家长互动等因素。这些因素的内涵、评价方法、标准等还需要进一步探索。

最后，本文的一个重要局限是由于每个教室只有一个观察员进行评分而无法计算评分者内部一致性指标（interrater agreement）。在未来的研究中，如果条件许可，我们将尝试两人评分，计算评分者内部一致性指标，严格监控评分质量。在条件有限的情况下，实施评分监测或多人评分，尽可能提高评分者内部信度（interrater reliability）[16]。

评价和研究学前教育质量对促进学前教育的发展有着重要的意义，它不仅可以对优秀的教师给予认可和嘉奖，还能将需要改进的地方及时反馈给教师和园长，并且能敦促我们对比标准，审视区域内学前教育整体的发展状况。因此有必要将此类评价和研究常规化、政策化。教育部近期发布的《3—6 岁儿童学习与发展指南》为评价学前教育的成果提供了标准和依据。对投入、过程和产出三方面的密切关注，定能大力促进我国学前教育质量的提高。

参考文献：

[1]OECD. Starting Strong III: A Quality Toolbox for Early Childhood Education and Care. Paris: OECD, 2012.

[2]刘占兰. 学前教育的普及必须兼顾质量的提高[J]. 学前教育研究, 2010, (10).

[3]OECD. Quality Matters in Early Childhood Education and Care: Finland. Paris: OECD, 2012.

[4]Cryer,D.,Tietze,W.,Burchinal,M.,Leal,T.,& Palacios,J. Predicting Process Quality from Strutural Quality in Preschool Programs: A Cross-Country Comparison. Early Childhood Research Quarterly,1999,14(3):339-361.

[5]项宗萍. 从“六省市幼教机构教育评价研究”看我国幼教机构教育过程中的问题与教育过程的评价取向[J]. 学前教育研究, 1995, (2).

[6]吴积静, 周洪飞. 幼儿园一日活动质量评价的探索[J]. 学前教育研究, 1994, (6).

[7]郭良菁. 上海市幼儿园教师专业发展自我评价体系研制简介[J]. 学前教育研究，2007, (07-08):10-13.

[8]虞永平. 幼儿园教学活动的评价[J]. 早期教育:教师版, 2005, (3).

[9]Wilson,M. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2005.

[10]Ishimine,K.;Tayler,C.; & Bennett, J. Quality and Early Childhood Education and Care: A Policy Initiative for the 21st century. International Journal of Child Care and Education Policy, 2010, 4(2):67-80.

[11]Stipek,D.,& Byler,P. The early childhood classroom observation measure. Early Childhood Research Quarterly, 2004, (19):375-397.

[12]Hamre,B.K.,Mashburn,A.J.,Pianta,R.C.,Lacasle-Crouch,J.,& La Paro,K.M. Classroom Assessment Scoring System Technical Appendix. 2006.

[13]Dunn,W.E.;Lantolf,J.P. Vygotsky's Zone of Proximal Development and Krashen's i+1: Incommensurable Constructs; Incommensurable Theories. Language Learning, 1998,48（3）:411–442.

[14]Wright,B.D., & Masters, G. N. *Rating scale analysis*. Chicago: MESA Press, 1982.

[15]Wu, M., Adams, R. J., & Wilson. ConQuest. Hawthorn, Australia: ACER Press,1998.

[16]Gwet,K.L. Computing Inter-rater Reliability and its Variance in the Presence of High Agreement. British Journal of Mathematical and Statistical Psychology, 2008,(61):29-48.

上期回顾

2012 年第 10 期（总第 74 期）

我国义务教育阶段教师培训的财政问题研究

摘要：本报告是“我国义务教育阶段教师培训机制研究”课题的部分成果。报告分析了中央及地方政府对义务教育教师培训的投入政策及投入现状，义务教育教师对培训费用的分担状况，据此对完善我国义务教育阶段教师培训的财政保障机制提出了初步的建议。

本《简报》由北京大学中国教育财政科学研究所主办；旨在反映本所最新的学术科研活动；相关内容仅体现作者本人观点，并不必然代表本所的立场。

简报内容仅供参考，如需转载须事先征得本研究所同意。

本期印发：1800 份

下载网址：http://ciefr.pku.edu.cn

主办单位：北京大学中国教育财政科学研究所

电子信箱：newspaper@ciefr.pku.edu.cn

责任编辑：毕建宏

电　　话：010-6275-9700

地　　址：北京市海淀区颐和园路 5 号
北京大学教育学院楼 413 室（100871）

传　　真：010-6275-6183