Doi：doi.org/10.70693/rwsk.v1i1.25

Rasch 模型在国内语言测评领域应用现状及展望

刘强[1]

（四川外国语大学，重庆 沙坪坝 400031）

摘要：Rasch 模型作为项目反应理论的经典模型，因结果可信度高、解释力强，在语言测试领域受到广泛关注。本文基于文献分析法，以中国知网（CNKI）2003—2023 年间 130 篇核心文献为样本，借助 Citespace 工具探究 Rasch 模型在国内语言测评领域的应用现状、问题与未来方向。研究发现：Rasch 模型的应用集中于信效度验证、评分一致性分析及分数等值研究，且多聚焦高等教育领域（占比 70%）。现存问题包括基础教育阶段应用不足（仅占 23%）、信效度验证方法单一化等。未来应推动 Rasch 模型在基础教育常态化测评中的应用，结合多模态效度验证方法，如访谈、认知诊断等，并探索其与计算机自适应语言测试，如 AI 命题与评估的融合路径。

关键词： Rasch 模型；语言测评；信效度；分数等值；计算机自适应测试

一、引言

在英语作为全球通用语背景下，各国更加重视英语教育。语言测试是推动语言教育取得更高质量的关键手段之一，语言测试不仅是为了检测语言教育的质量或语言学习者的学习成效，同时也在学习过程中提供形成性评价，为语言学习者提供反馈，促进其学习。

George Rasch（丹麦数学家）针对其 CTT（Classical Test Theory）经典测量理论的不足，于是构思出了单参数 IRT（Item Response Theory）模型，即 Rasch 理论。Rasch 模型区别于其他模型最重要的一点就是 Rasch 模型需要所得到的数据与自身模型想相符合。只有数据与模型符合时才能有良好的吻合度（model-data fit）才能得到有意义且相对客观的量尺。此处的拟合度分为两个方面，被试者拟合度（person fit）和项目拟合度（item fit）。关于拟合度的计算，是由被试在每道题的真实反应（观察值），和模型自身计算得到的期望值进行对比，两者残差越大即越不吻合。其次，在 Rasch 模型中，本模型假设题目难度与被试能力相匹配。

自 Rasch 模型由陈国富等人首先在国内进行了模型的参数研究（陈国富，1987）以来，该模型就广泛应用于各大学科研究领域之中，如外国语言文学、教育学、心理学、医学等领域。其中，外国语言文学占比最大，且主要以语言测评应用研究为主。由此，作者将对分析 Rasch 模型在国内语言测评领域研究现状，总结特点及其趋势，以期为相关研究者提供研究方向与建议。

二、研究问题与方法

（一）研究问题

[1] **基金项目**：四川外国语大学研究生科研创新项目(SISU2024XK120)

作者简介：刘强(2000—)，男，硕士研究生，研究方向语言教育、语言测试；

Rasch 模型作为语言测试的重要辅助工具，其对 Rasch 模型的拟合程度可对试题的信度、效度、以及评分信度等具有重要参考意义。本研究分析并梳理国内 Rasch 模型相关文献，以期解决以下研究问题：

1．近年来国内 Rasch 模型在语言测试领域的研究总体发文趋势如何？

2．研究现状如何？存在什么问题？未来研究方向是什么？

（二）研究方法

1．文献收集

为回答以上问题，本研究在 CNKI（中国知网）进行检索，以“Rasch”为主题进行检索，共检索到 373 篇期刊文献，由于 Rasch 模型在心理学、教育学、医学等诸多领域都有所应用，故本研究作者按引用率排序剔除与语言测试不相关文献后剩余 130 篇文献。

2．数据分析

为探究第一个问题，本研究利用 Citespace 对 130 篇文献进行分析，分析设置年份为（2003-2023），通过其历年发文量探求 Rasch 模型在语言测试领域的总体发文趋势。为探究第二个问题，本研究将通过 citespace 分析这 130 篇文献，产出其关键词中心度图、关键词共现图、关键词聚类图和关键词时间线图，并对其观察与分析。

三、结果与分析

（一）总体发文趋势

如图 1，通过对其文献分析，关于 Rasch 模型自系统性的引入国内以来（ 罗冠中，1992），真正开始应用在国内语言测试领域是在二十一世纪以后。最早的应用在 Rasch 模型开始于四六级考试等值应用研究（朱正才,杨惠中,杨浩然，2003），随后国内学者纷纷开始注意到此模型

图 1

的应用，从本世纪初及其后十年，该模型在语言测试界研究发文量呈现上升趋势。随后发文量总体呈现波动中下降的趋势，说明对 Rasch 在语言测试领域应用的研究热点有所下降。如图 2，关键词中心度前五名由高到低分别是：信度、效度、效度验证、概化理论和口试。可知，Rasch 模式在语言测试中，主要集中于信效度验证等领域，特别是在主观测试口试中。

图 2

频次	中心度	年份	关键词
12	0.16	2008	信度
11	0.14	2010	效度
11	0.13	2010	效度验证
6	0.03	2008	概化理论
4	0.09	2016	口试
4	0	2016	语言测试
3	0.08	2013	质量分析
3	0.04	2007	评分量表
3	0.03	2010	偏差分析
3	0	2013	评分效度

（二）研究现状

通过对其文献分析，从关键词时间线图（图 4）国内将 Rasch 应用于语言测试领域最开始始于四六级考试，从关键词聚类图（图 3）中最突出的关键词主要有：效度验证、效度、信度、口试、评分量表、概化理论等，由此可知 Rasch 在语言测评领域研究最大的特点主要是进行信效度验证。在关键词聚类分析（图 3）中，Q 值=0.803，S 值=0.949，说明其分析结果具有有效性。分析结果显示，有效聚类主要有四类，即：#0 效度、#1 效度验证、#2 大学英语、#3 评分信度和#7 多面 Rasch 分析，由于#1 与#2 基本属于一类研究，故总聚类主要有四类：#0 效度、#2 大学英语、#3 评分信度和#7 多面 Rasch 分析。在#2 大学英语关键词聚类下，分级考试、等

图 3

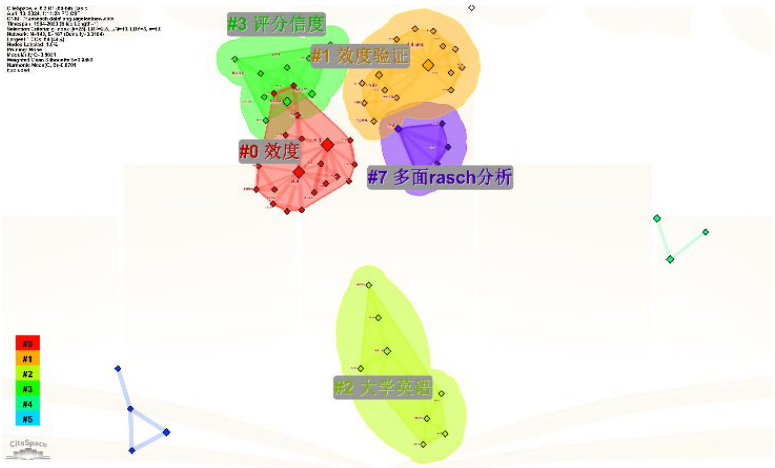
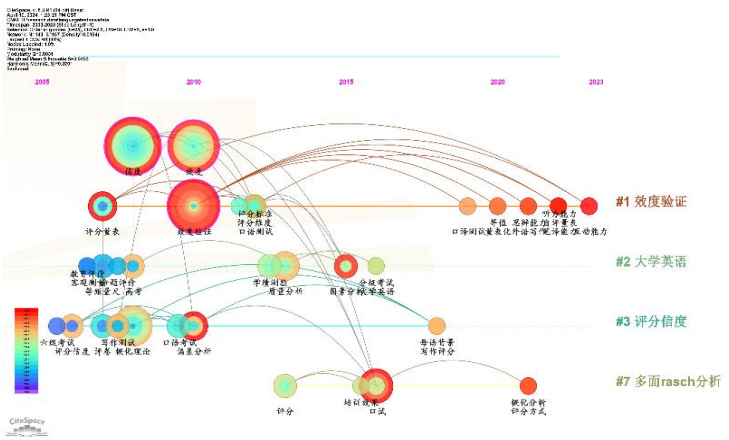


图 4



距两尺、分数等值等相关关键词较为突出，因此，在高等教育阶段中，关于分数等值研究也是 Rasch 模型应用的一大热点。结以上信息，本研究认为 Rasch 模型应用在语言测试领域主要包含三个方面：测试工具、量表、考生的信度研究；测试工具效度研究；分数等值研究三个方面。

1．信度研究。

语言测试的信度主要指的是某一测试结果的可靠程度，即某一语言测试的结果在多大程度上反映了被试的语言能力水平。而影响测试信度的因素较多。首先，对于测试工具本身来说主要是受到被试样本量和区分度的影响，被试样本量是否具有代表性、被试样本量是否足够多、试题区分度是否大，信度与这些维度呈正相关。其次，在组织测试过程中，测试环境，如考场温度、噪音、光线；监考人员是否和善等这些

因素都会影响考生答题情况，进一步影响测试效度。最后，在测试后评分员的宽严程度、前后一致性也会影响测试信度。

（1） 测试工具的信度研究

国内对于测试工具的信度研究主要涉及在对于语言测试题和量表的信度研究。对于试题研究，如（2011，王立非）探讨了全国国际商务英语考试(二级)能力量表的设计,以 2009 年的二级考试数据为依据,对考试的信度和效度进行分析,运用多层面 Rasch 模型,测量了试卷各题目的难度,以检验该考试的区分度和权威性。统计结果显示,该考试的信度、效度以及难度都达到了测试学的要求,具有较好的稳定性,适用于大规模考试推广。对于量表研究，如（杨建原，曾薇，2012）以贵州师范大学共 90 个大一新生在症状自评量表上的数据为实例,讨论 Rasch 等级模型在设计和修订等级量表中的应用,以及如何应用 Rasch 等级模型的某些参数如选项频率、平均测量值、临界值、概率曲线、选项拟合指数等来对等级量表的选项分类数目、选项标签进行直观的分析和检验,从而获得高质量量表。（史剑雄，王佶旻，2022）聚焦构建汉语作为二语的诊断性写作测试评分量表。多面 Rasch 模型检验结果表明：评分量表能有效地区分学生的写作能力；评分员内在一致性良好；量表的不同维度的难度存在显著差异；各维度的分值设置合理，分数使用不存在无序性。整体而言，评分量表的质量是令人满意的，通过访谈也得到了积极的反馈。总之，Rasch 模型不仅量化了测试工具的稳定性，更通过参数动态反馈机制，解决了传统信度检验中“一刀切”的局限性。其多维度拟合分析（如题目难度、考生能力、评分标准）为大规模考试的标准化推广提供了数据驱动的决策依据，尤其在跨区域、跨群体测试中展现出独特的适配性。

（2）评分信度研究

语言测试信度极大的受到评分人员的影响，包括评分人员对于评分标准的解读，评分人员的宽严程度以及前后一致性。如（温倩，2019）对商务谈判口译行为测试进行效度验证，对于评分人员信度部分采用了多层面 Rasch 模型进行分析，结果显示评分员表现出良好的评分员间一致性。（吴雪峰，2018）探索构建写作评分标准的一般规律,尝试设计英语写作测试评分标准模型，结果表明:评分标准区分度和效度较好,评分员与评分标准间存在显著偏性交互作用;个别分数段的使用存在非拟合现象。Rasch 模型突破了传统“评分员培训”的经验主义范式，通过残差分析和偏差指数（如 Infit/Outfit），将主观评分行为转化为可量化、可干预的客观指标。这不仅为动态监控评分质量提供了实时工具，更推动了评分标准从“静态描述”向“动态校准”的范式转型。

2.效度研究

语言测试的效度主要指的是一种相关性，即测试与预期测试的目标是否达成一致，换句话说即该测试是否考察到命题人预期考察到的内容与能力(Lado,1961)。测试界认为的效度主要包括结构效度(construct validity）、内容效度（content validity）、同期效度（concurrent validity）、预测效度（predictive validity）和表面效度（face validity）。Rasch 模型主要是对同期效度和预测效度进行检验，即测试结果与外在标准的的一致性，也称标准参照效度。相关研究也主要分为语言测试题效度、量表效度和评分效度研究。

（1）语言测试题的效度研究

对于语言测试题的效度研究的核心思想是，该测试题是否考察到应该开叉的能力，或者该语言测试题是否有利于语言的学习。如：运用 Rasch 模型对概要写作作为一项综合技能考试进行效度验证，结果表明评分量表能够较好地区分考生的能力，但仍需进一步改进；评分单项中语言使用(LU)难度最大，原文信息使用(SU)难度最小（李久亮，2022）。Rasch 模型通过概率模型（如题目特征曲线）将抽象的“效度”概念转化为可操作的“拟合度”指标，为测试题的构念效度验证提供了数学框架。其核心贡献在于将效度研究从“相关性统计”提升至“因果机制分析”，例如通过题目难度参数识别能力培养的瓶颈点。但当前研究对“表面效度”（face validity）和“社会效度”（如测试对教学的反拨效应）的关注不足，导致效度验证偏重技术理性，忽视教育生态的整体性。

（2）量表效度研究

对于语言测试界的量表效度主要集中于该量表是否能反映学生的真实语言水平。如：采用多层面 Rasch 模型对配对口语测评评分量表进行效度验证，研究发现总评分量表与互动能力评分量表可以有效的区分不同口语能力和互动能力的考生（邹润,2023）。Rasch 模型通过“分离度指数”（separation index）和“信噪比”（person/item reliability），实现了量表效度从“整体评价”到“微观诊断”的跨越。其核心优势在于通过概率模型揭示量表的隐性结构（如维度权重、选项阈值），为个性化反馈提供了数据基础。

（3）评分效度研究

评分效度即对于测试结果能否有效的说明考生的真实语言水平。如：（杨志强，许吟雪，2018）用多层面 Rasch 模型,通过分析 PRETCO 口试的评分结果以探究其评分效度。研究发现 PRETCO 口试评分效度较高,其评分结果能够有效区分考生的口语水平,评分员评分的自身一致性总体较好。研究同时发现 PRETCO 口试评分存在以下问题:评分员的宽严度差别显著,个别评分员的内部一致性较差;少数评分员和考生的交互作用存在显著差异;评分员和四项任务之间也出现了不同程度的偏差。Rasch 模型通过交互效应分析（如 Facet 交互项），将评分效度从“静态一致性”扩展至“动态情境适应性”，为高复杂度评分任务（如口译、写作）的效度验证提供了方法论创新。其突破性在于将评分效度与任务设计、评分员认知特征联动分析，推动了效度研究从“结果验证”到“过程优化”的转型。

3.分数等值研究

Rasch 模型的一大特点在于可以将考生能力与分数放在同一量尺，这样的特点对于国内大规模考试（如 CET4,CET6,TEM-4TEM8 等）的分数等值换算具有重要意义，像此类大规模通过性考试，如若直接将原始分数作为参考，那每年的通过率将会起伏很大，因此为保障考试公平性，将分数与能力对等，通过等值运算得出考生报告分。如：（朱正才，2005）对现有的大学英语四、六级考试分数等值模式中存在的若干问题进行了深入的分析,并提出了新的解决方案，一个基于铆题设计和两参数 IRT 模型的解决方案。（朱正才,杨惠中,杨浩然，2003）Rasch 模型在 CET 考试分数等值中的应用，描述了 Rasch 模型在中国大学英语四、六级考试的分数等值系统中的应用情况,并使用真实考试数据进行了 Rasch 模型背景下的分数等值实验,对分数等值过程中的诸多问题进行详尽的分析和探讨。Rasch 模型在分数等值中的核心价值在于构建了“题目无关”的能力量尺（θ值），使不同批次、不同形式的考试成绩具备可比性。其方法论革命性体现在将等值从“统计调整”升级为“测量理论驱动”的标准化流程，尤其在高风险考试（如高考、职业认证）中保障了选拔公平性。

四、存在问题

（一）Rasch 模型在基础教育语言测试应用不足

在本研究 130 篇文献中，将 Rasch 应用于基础教育阶段语言测试的文献只有 30 余篇。占总发文数量的 23%。而在高等教育阶段则有 90 余篇，占总文献数量的 70%。Rasch 应用在大学英语较多而基础教育阶段语言测试较少的原因主要有以下几点。1)信效度验证需求不同。高等教育对其测试工具和评分人员的信效度验证需求更多，而基础教育语言测试主要集中在大规模高风险考试为主，教师对其测试工具和人员的信效度检验要求较低。2）测试题型不同。基础教育阶段除写作外，基本是客观测试为主，而大学英语测试有其更多的主观测试，如：口译测试、口语测试等，这就需要多层面 Rasch 模型对其信效度验证。并且，现有研究多聚焦高利害考试（如 CET、TEM），对校本测评的关注不足，且缺乏长期追踪数据验证工具信度的时效性。

（二）信效度验证方法单一

对于 Rasch 模型应用于语言测试领域而言，对其信效度的分析多只是对其结果与模型的拟合程度进行探讨。这既是 Rasch 模型的特点、优点，同时也是其缺点。特别是对于效度验证而言，效度的核心是是否考察到应考察的内容与能力。这对于终结性评价确实是一种有参考意义的分析方法，但是对于过程性评价而言，测试评价的目的不是测试，而在于促进被试的学习，在这个过程中，单一的 Rasch 模型分析来试卷的信效度是不充分的。总之，现有研究多依赖横截面数据，未能结合认知心理学方法（如眼动追踪、决策过程建模）深入解析评分偏差的认知根源，限制了模型的解释深度。

总体而言，Rasch 模型在国内语言测评领域的应用呈现“高教主导、基教薄弱”的失衡格局。这一现状可能源于基础教育阶段对测评科学性的重视不足，以及客观题主导的测试形式与 Rasch 模型适配性较低。此外，过度依赖模型拟合度检验而忽视多维度效度证据，进一步限制了其在形成性评价中的应用潜力。

五、未来展望

（一）基础教育阶段测试模型应用常态化

以往研究多集中于大规模考试，如四六级、专四专八、中高考。而对于校本命题，基础教育阶段语言类教师通常对于考试测评领域相关专业知识是缺乏的，但是仍然需要学校英语命题工作，因此命题质量得不到保障，对于测试试题的难度与学生匹配度的把握、试题区分度、测试信效度方面没有统一的检测方法。而 Rasch 模型由于其自身特性，则可以对上述要素进行检验，对校本、班本测试的信效度也能进行分析，

提升校本命题质量。

（二）测试效度多种方法结合验证

效度的核心思想在于测试是否测试了预期考察的内容与能力。而 Rasch 模型多只是作为工具对测试结果与能力匹配度进行分析。但是某些能力并不能通过一次测试就能完全测出，如：阅读能力考察的推理判断能力等，这些能力的运用具有开放性，并不是一次测试结果就能证实，我们还应该考虑访谈、有声思维、观察法多种方法相结合。

（三）模型与计算机自适应语言学习相结合

Rasch 模型作为一种基于概率的潜在特质测量模型，在心理测量与教育评估领域具有深厚理论基础，其与人工智能及计算机领域的结合将在数据处理、模型优化和应用场景扩展等方面开启新的可能性。在技术融合层面，Rasch 模型的参数估计过程可通过深度学习框架实现加速与优化，例如将项目难度、被试能力等参数的联合估计转化为神经网络权重的学习问题，利用反向传播算法处理大规模测试数据，同时结合贝叶斯方法增强模型对稀疏数据的适应性。在应用场景拓展方面，Rasch 模型可与自然语言处理技术结合，构建智能测评系统，通过分析学习者对开放式问题的文本响应，动态校准认知能力与题目特征的关系；在自适应学习系统中，模型可嵌入强化学习框架，根据实时交互数据动态调整题目推送策略，实现个性化的学习路径优化。此外，在计算机视觉领域，Rasch 模型可辅助评估图像标注任务的难度层级，为 AI 训练数据的质量控制提供量化标准。值得注意的是，这种融合需突破传统 Rasch 模型的线性假设限制，通过引入非线性扩展模型或与深度生成模型结合，处理复杂的高维交互数据。同时，在可解释性层面，需平衡 AI 黑箱特性与测量模型的透明性要求，开发可视化工具揭示 AI-Rasch 混合模型的决策逻辑。未来，随着边缘计算与联邦学习的发展，分布式 Rasch 模型可能实现跨设备、跨平台的协同能力评估，在保护数据隐私的前提下提升模型的泛化能力，为教育公平、人才选拔等社会议题提供智能化解决方案。

参考文献

[1]陈富国,李伟明.Rasch 模型及其参数的近似估计[J].江西师范大学学报,1987(02):75-81.
[2]罗冠中.Rasch 模型及其发展[J].教育研究与实验,1992(02):40-43.
[3]朱正才,杨惠中,杨浩然.Rasch 模型在 CET 考试分数等值中的应用[J].现代外语,2003(01):69-75.
[4]王立非,江进林.全国商务英语考试的设计与信效度研究[J].外语与外语教学,2011(06):35-40.
[5]杨建原,曾薇.Rasch 模型在等级量表设计中的应用[J].中国考试,2012(05):12-18.[6]史剑雄,王佶旻.诊断性写作测试评分量表的构建研究[J].语言文字应用,2022(03):114-123.
[7]温倩.多面 Rasch 模型的商务谈判口译行为测试效度验证[J].中国外语,2019,16(03):73-82.
[8]吴雪峰.英语概要写作评分量表研制的实证研究[J].中国外语教育,2018,11(02):57-66+86.
[9]Lado.R. L anguage Testing：The Construction and Use of Foreign L anguage Testing. London：Longman，1961.
[10]李久亮.基于多层面 Rasch 模型的概要写作效度验证[J].外语测试与教学,2022(03):38-48.
[11]邹润.基于 Rasch 模型的配对口语测试评分量表的效度研究[J].解放军外国语学院学报,2023,46(05):20-29.
[12]杨志强,许吟雪,全冬.PRETCO 口试评分效度研究[J].重庆三峡学院学报,2018,34(02):121-128.
[13]胡壮麟.ChatGPT 谈外语教学[J].中国外语,2023,20(03):1+12-15.DOI:10.13564/j.cnki.issn.1672-9382.2023.03.003.
[14]何莲珍,张洁.多层面 Rasch 模型下大学英语四、六级考试口语考试(CET-SET)信度研究[J].现代外语,2008(04):388-398+437.
[15]吕晓轩,任伟.基于 Rasch 模型的《中国英语能力等级量表》笔译自评量表效度研究[J].外语教学,2022,43(01):57-61+94.

Current Status and Prospects of the Rasch Model in Language Assessment in China

Qiang Liu

[1]（*Sichuan International Studies University,Chongqing,400031*）

Abstract:
As a classical model within Item Response Theory (IRT), the Rasch model has gained significant attention in the field of language assessment due to its high reliability and strong explanatory power. This study adopts a literature analysis approach, using 130 core articles from the China National Knowledge Infrastructure (CNKI) database published between 2003 and 2023, and employs CiteSpace to examine the current status, existing issues, and future directions of Rasch model applications in the Chinese context of language testing. The findings indicate that Rasch model applications are predominantly focused on the validation of reliability and validity, rating consistency analysis, and score equating, with approximately 70% of studies concentrating on higher education contexts. However, its use in primary and secondary education remains limited (only 23%), and the methods for validity and reliability assessment are relatively homogeneous. Future research should promote the routine use of Rasch modeling in basic education assessments, integrate multimodal validity verification methods—such as interviews and cognitive diagnostic assessment—and explore its convergence with computer-adaptive language testing technologies, including AI-assisted test item generation and automated evaluation.

Keywords: Rasch model; language assessment; reliability and validity; score equating; computer-adaptive testing