 DOI: https://doi.org/10.12345/jxffcxysj.v8i6.24811

Validity of Test Score Interpretation in Argument-Based Models

Xu Xiaoyan

Inner Mongolia Institute of Educational Research and Evaluation，Inner Mongolia，Hohhot，010011

Abstract

Validity, as a core indicator in educational measurement, has evolved through four developmental stages of validation models: criterion validity, classification validity, evidence integration, and systematic argumentation. This progression reflects a shift from single-dimensional technical verification to multi-dimensional logical reasoning. This paper focuses on the Interpretation/Use Argument (IUA) model proposed by Kane, systematically examining its theoretical framework and practical significance. The IUA model incorporates Toulmin's practical argumentation framework, reconstructing validity validation into a dynamic logical chain composed of "scoring," "generalization," "extrapolation," and "decision-making." It emphasizes ensuring the rationality of test score interpretation and use through multi-level reasoning and rebuttal mechanisms. This model transcends the limitations of traditional validity validation, shifting research from "proving tool effectiveness" to "arguing interpretative rationality." By integrating multi-source evidence, addressing decision consequences, and adopting dynamic updating mechanisms, it enhances the social trust and fairness of assessments. Case studies demonstrate the successful application of the IUA framework in high-stakes assessments such as the TOEFL and Physician Licensing Examinations, driving the transformation of validity validation standards toward practical rationality. Future research should explore intelligent technology empowerment and cross-cultural validity argumentation to address challenges in complex assessment ecosystems.

Keywords

Validity ; Validity validation; IUA

论证模式下考试分数解释效度研究

徐霄雁

内蒙古自治区教育科学研究与监测评估院，中国·内蒙古 呼和浩特 010011

摘　要

效度作为教育测量的核心指标，其验证模式历经效标效度、分类效度、证据整合及系统论证四个发展阶段，逐步从单一技术验证转向多维度逻辑推理。本文聚焦Kane提出的解释使用论证模式（IUA），系统梳理其理论架构与实践价值。IUA模式引入图尔敏实用论证框架，将效度验证重构为由“评分”“概化”“外推”“决策”等环节组成的动态逻辑链条，强调通过多层级推理与反驳机制确保考试分数解释与使用的合理性。该模式突破传统效度验证的局限性，推动效度研究从“工具有效性证明”转向“解释合理性”论证，并通过整合多源证据、关注决策后果及动态更新机制，提升考试的社会信任与公平性。案例分析表明，IUA框架已成果应用于托福考试、医师资格考试等高风险测评，推动效度验证标准向实践理性转型。未来研究需进一步探索智能技术赋能、跨文化效度论证等方向，以应对复杂测评生态的挑战。

关键词

考试分数解释效度；效度验证；解释使用论证模式

效度作为教育测量的核心指标，直接决定考试分数的解释与使用的科学性和有效性。随着教育与心理测量理论的演进，效度研究从单一效标验证逐步发展为基于多维度证据的系统论证。随着教育与心理测量理论的发展，关于效度研究的理念、验证模式也随之不断发展。历经 4 个发展时期，效度验证模式的整体性和系统性日益提高。本文梳理效度概念及验证模式的发展脉络，重点分析 Kane 提出的解释使用论证模式（Interpretation/Use Argument, IUA）的理论创新与实践意义。

1 效度概念的历史演变

效度不仅是衡量一个考试项目质量高低的重要指标，而且是指导考试设计、题目编写、试卷编排、考试实施、赋分、分数报告与分数使用等工作的根本依据。随着教育与心理测量理论的发展，关于效度研究的观念也随之不断发展。1954 年《关于心理测验和诊断的技术建议》首次将效度分为预测效度、共时效度、内容效度与结构效度。这一阶段强

【基金项目】内蒙古自治区2022年度教育考试招生研究专项课题重点项目《论证模式下考试分数解释效度研究》（项目编号：KSZX202228）。

【作者简介】徐霄雁（1988–），女，中国内蒙古呼和浩特人，硕士，助理研究员，从事教育评价与监测研究。

调通过统计方法（如相关系数）验证测验与效标的关系，但存在效标选择的主观性和局限性问题。1988 年 Messick 提出整体效度观，将效度定义为“测验分数解释与使用的合理性程度”，强调证据来自多方面（如内容、结构、结果等）。1999 年和 2014 年版的《教育与心理测验标准》把效度定义为“证据及理论对测试分数解释与使用的支持程度”。从效度概念的发展可以看出，教育和心理测量学者们关于考试效度的看法越来越谨慎，他们不再一般地谈论一个考试的效度，而是论证将一个考试应用于某一特定目的时候，某一次考试的分数解释的效度。学界逐渐明确效度验证的对象并非测验本身，而是“分数解释与使用的合理性”。测试质量（如信度）是效度的前提，但效度关注的是分数应用的逻辑支持，而非工具属性。

2 效度验证模式的发展：从证据积累到系统论证

进入效度证据整合时期，学界开始突破传统分类的桎梏。1989 年，Messick 在《教育测量》中提出“整体效度观”，强调效度并非单一指标的叠加，而是需通过实证数据与理论逻辑的协同作用，综合评估分数解释的科学性与应用决策的合理性。他构建的分层效度框架以构念效度为核心，整合内容、过程、结构等多维度证据，推动效度研究从“碎片化验证”转向“系统性论证”。这一理念在 1999 年修订的《教育与心理测验标准》中进一步深化，明确效度支持需涵盖五大证据类型：内容覆盖的充分性（如试题与目标领域的一致性）、反应过程的规范性（如考生答题行为的可解释性）、内部结构的稳定性（如因子分析结果与理论模型匹配度）、外部变量的关联性（如分数与效标间的预测效度）以及应用结果的实效性（如考试决策对教育公平的影响）。至基于论证的效度验证时期，Kane 于 2006 年引入图尔敏的实用论证模型，提出解释使用论证模式（IUA），彻底革新效度验证的范式。该模式将效度评估重构为动态逻辑推理过程，包含六大要素：资料（D）（如考生原始作答）、必要条件（B）（如评分规则的科学性）、理由（W）（如构念理论支撑）、限定（Q）（如考试情境的适用范围）、反驳（R）如评分偏差风险）和结论（C）（如分数解释的有效性）。其核心在于通过逐层递进的论证链条，系统检验从评分到决策各环节的合理性。例如，在“评分推断”阶段需验证评分者一致性，而“概化推断”则需证明试题样本对目标能力域的代表性。这一模式不仅规避了传统效标验证的片面性，更通过引入“反驳机制”，要求研究者主动识别并回应效度威胁，从而增强论证的严谨性与透明度。效度论证需满足：清晰性（逻辑框架明确）、完整性（推论链条无断裂）、可接受性（假设合理）三大标准，确保论证过程兼具科学性与说服力。

3 Kane 解释使用论证模式

Kane 首先将 Toulmin 的论证框架引入了效度验证，引入后的效度论证框架包含资料、支撑、理由、限定、例外和主张 6 个基本要素。每个推断从一个资料开始，以一个主张结束，运用从相关理论或实证研究中收集的支撑或证据证明理由，接受例外的限定后，推导出主张（图 1 是效度论证基本过程）。

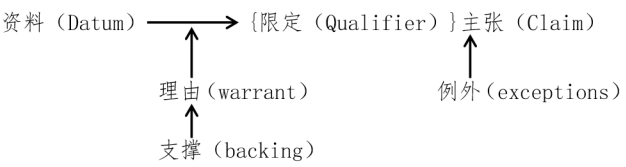

图 1 效度论证基本过程

Kane 确定了两种类型的论证：解释性论证，然后是效度论证。他后来扩展了该框架，在解释性论证中增加了考试使用推论 (test use)，将解释性论证扩展为解释 / 使用论证（interpretation/use argument，IUA）。Kane 解释说，验证是一个持续的过程，该过程：首先概述解释性论证中考试评分的拟定解释和使用，从概念上将考试性能与基于考试评分的主张和决定联系起来；然后在效度论证中评价这些拟定解释的合理性。换言之，解释性论证概述了验证需要经过的步骤以及以何种方式进行，而效度论证是指证据支持或质疑解释性论证的程度。基于论证的效度验证模式的特征可总结如下：（1）研究者指定各种分数的解释和用途；（2）基于考试分数提出的主张或推断被用于构建解释性论证；（3）研究者使用解释性论证作为收集证据构建效度论证的框架。该方法为验证研究者如何构建效度论证提供了系统化的过程，并允许研究者或考试开发者根据考试分数和支持分数的解释和使用所需的证据类型，灵活地确定他们想要提出的主张。

Kane 认为效度验证贯穿于从施测到决策的整个考试过程中：（1) 评分 (scoring/evaluation)，用观察到的考生表现推断观察分；（2) 概化 (generalization)，利用观察分估计考生在平行任务和测试中所期望的分数；（3) 外推 (extrapolation)，用观察分推断全域分。效度研究不再是简单的计算考试分数与效标之间的相关系数，也不再是简单地收集证据或事实，而是一个持续的、层层深化的、多层级的研究过程。遵循 Toulmin 论证模式，前一推论中由“理由”所推导出的“主张”，可以成为下一个推论的“理由”，每步推论紧密相连，形成一个完整的从起点资料到最终主张的论证系统。

基于论证的效度验证框架已经被许多语言测试开发者和研究者用来验证高风险和低风险测试。其中，新托福考试是应用论证模型开展效度论证的典范，Chapelle 等人在 Kane 的基础上，进行了六个步骤推论：（1）领域描述，（2）评价，（3）概化，（4）解释，（5）外推，（6）使用（图 2 是新托福效度论证框架）。

图 2 新托福效度论证框架

4 Kane 解释使用论证模式的地位与意义

相较于传统效度验证方法，IUA 模式实现三大突破：在方法论层面，将统计验证转变为逻辑论证，要求研究者构建可辩驳的推理链条；在认识论层面，强调效度是暂时性结论而非终极判断，加拿大医师资格考试（MCCQE）每五年更新效度论证便是明证；在价值论层面，将决策后果纳入效度范畴。

Kane 模式将效度验证视为“支持者与反对者的辩论”，通过动态论证过程界定考试的应用边界。其核心贡献在于：（1）逻辑严谨性：引入图尔敏模型，构建多层级推理网络，避免传统效标验证的片面性；（2）系统性：覆盖评分、概化、外推、决策全流程，形成闭环验证；（3）实践导向：指导 ETS、剑桥英语等机构优化效度研究，如托福考试采用 IUA 框架整合证据链。其对于效度验证的实践意义在于：（1）规避分数误用：通过限定条件（Q）和反驳（R）明确分数解释的适用范围；（2）提升社会信任：公开论证过程增强考试透明度，维护考试公平性；（3）推动范式转型：从“证明测验有效”转向“论证解释合理”，契合复杂心理属性的测量需求。

该模式推动效度验证标准从技术理性转向实践理性。美国教育研究协会（AERA）2014 版《教育与心理测试标准》明确要求效度论证包含决策后果评估，英国剑桥考评集团（Cambridge Assessment）建立论证跟踪系统实时监测考试决策的社会影响，这些实践变革均体现 Kane 模式的理论渗透。

效度研究从单一效标迈向系统论证，反映了教育测量学对人本复杂性的深刻认知。解释使用论证模式将效度验证转变为持续的社会协商过程，这种动态认知框架恰适应当代测评生态的复杂性。Kane 的解释使用论证模式通过结构化推理链条，将效度验证提升为科学辩论过程，不仅革新了方法论，更重塑了考试公平性与有效性的实现路径。未来研究需在技术赋能与跨学科合作中深化探索，进一步突破效度验证的实践瓶颈，使效度理论始终保持对测评实践的解释力与指导性，推动教育测评的可持续发展。

参考文献

[1] 周群.基于论证的我国高考开发质量评价模型研究[D].华东师范大学,2011.

[2] 杨志明, 林兰兰. 基于效度证据的英语测验研发[J]. 教育测量与评价, 2021, (08): 3-9.

[3] 肖媛, 李玲玉, 李群锋, 张欣, 李佩泽. 基于证据观的医学汉语水平考试（MCT）效度研究[J]. 天津师范大学学报(社会科学版), 2021, (04): 52-57.

[4] Kane, M. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527–535.

[5] Kane, M. (2013). Validating the interpretation and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

[6] Kane, M. (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice. 23(2), 198–211.

[7] Kane, M. (2021). Articulating a validty argument. In G. Fulcher & L. Harding (Eds.), The Routledge Handbook of Language Testing (pp.34-47).

[8] 周艳琼.《中国英语能力等级量表》自评量表的效度验证[J].现代外语,2021,44(01):101-112.

[9] 谢小庆. 效度:从分数的合理解释到可接受解释[J]. 中国考试, 2013, (07): 3-8.

[10] Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Building a validity argument for the Test of English as a Foreign Language.

[11] Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Language Testing, 27(4), 443–469.

[12] Ching-Ni Hsieh.(2024).Building a Validity Argument for the TOEFL Junior Tests.TOEFL® Research Report.