



国际中文教育通讯
第4卷，第1期，19-32
<https://doi.org/10.46451/iclec.260812>

收稿：2026年6月5日
接受：2026年8月10日
出版：2026年8月17日

HSK2.0 与 HSK3.0 书写测试任务特征与潜在构念对比研究

李蓉
江汉大学，中国
(电邮: lirongjhun@163.com)

摘要

针对 HSK 书写部分合格率偏低及构念测量偏窄等问题，本研究基于构念效度理论与交际语言能力模型，从构念代表性、任务内在关联性及难度梯度适配性三个维度，系统对比 HSK2.0 与 HSK3.0，并辅以实证三角互证。研究发现：HSK2.0 的测查重心较多倾斜于低阶形式与认知技能，构念覆盖相对偏窄；HSK3.0 则实现了语篇、语用与策略构念的全维度覆盖，任务设计向高阶认知与深层功能关联跃迁，难度梯度更契合能力发展。同时，实证调查揭示 HSK3.0 在实测中面临高阶计分信度存疑、四至五级字数跨度易致认知超载及学术写作缺位等潜在挑战。本文据此提出警惕应试反拨、平缓认知跨度及拓展测评边界的优化建议。

关键词

HSK2.0, HSK3.0, 书写部分, 构念效度, 交际语言能力

A Comparative Study of Task Characteristics and Potential Constructs in the Writing Sections of HSK 2.0 and HSK 3.0

Rong Li
Jiangnan University, China

Abstract

Addressing the low pass rate and narrow construct coverage in the HSK writing section, this study systematically compares HSK 2.0 and 3.0 based on construct validity theory and the Communicative Language Ability model. The comparison focuses on three dimensions: construct representation, internal task relevance, and difficulty gradient alignment, complemented by empirical triangulation. Results show that HSK 2.0 tends to focus more heavily on lower-order forms and cognitive skills, with relatively narrow construct coverage. In contrast, HSK 3.0 achieves full-dimensional coverage encompassing discourse, pragmatics,

and strategy constructs. Its task design shifts toward higher-order cognitive skills and deep functional relevance, featuring a difficulty gradient more consistent with capability development. Meanwhile, empirical findings reveal potential challenges for HSK 3.0: uncertain scoring reliability for high-level constructs, potential cognitive overload from the steep word-count increase between Levels 4 and 5, and the absence of academic writing. Consequently, this study proposes optimization suggestions: preventing negative washback, smoothing the cognitive transition, and expanding assessment boundaries.

Keywords

HSK 2.0, HSK 3.0, writing section, construct validity, communicative language ability

引言

汉语水平考试（HSK）是国际中文教育领域使用范围最广、影响力最大的标准化语言水平测试，其测试内容、题型设计与质量直接关系到全球中文学习者的能力评价、学习导向与教学实践。在当代大规模国际语言测试发展的宏观图景中，HSK 考试设计正经历着从早期的“结构主义（侧重语言形式）”向“交际主义（侧重真实语境下的语言功能与策略运用）”的深刻转型。为持续提升考试的科学性与适用性，深化对 HSK 考试改革的学理审视显得尤为关键。

从发展历程来看，HSK 先后经历了三个重要阶段。20 世纪 90 年代初，标准化汉语水平考试（HSK 1.0）应运而生；2009 年，HSK 实施改革并发布新标准，正式进入 2.0 阶段，全球报考人数实现激增。然而，随着国际中文教育的快速发展，HSK 2.0 在考查内容上逐渐难以满足日益复杂的真实交际需求。直至 2021 年，《国际中文教育中文水平等级标准》（教育部、国家语言文字工作委员会，GF 0025-2021，以下简称《等级标准》）发布，将汉语水平划分为三等九级，为 HSK 新一轮改革提供了重要依据。经过数年筹备，HSK 3.0 于 2026 年 1 月正式试点开考，在考试结构与题型设计等方面实现了全面升级，其核心动因正是为了克服旧版本在交际构念覆盖上的局限。

在听、读、写三大模块中，书写部分因实用性强、综合度高，长期以来是 HSK 考试的重点与难点。丁安琪等（2022）基于 2018 年全球 HSK 考试数据分析指出，书写部分合格率整体偏低，是影响考生通过考试的重要现实瓶颈。这一教学与实测中的痛点，在客观上暴露出传统书写测试在任务设计与能力映射上的不足，进而迫切需要通过深化改革来予以破解。正是在此现实诉求与国际化交际转向的双重驱动下，HSK 3.0 的书写改革主动向国际主流语言测试看齐。以国际英语测试系统（IELTS，以下简称“雅思”）的学术类写作为例，其任务类型（如图表描述与话题论述）与 HSK 3.0 的中高等级高度相似，均强调对图表信息的概括、观点的逻辑论证以及语体的恰当运用。然而，两者的构念侧重仍有差异：雅思在评估标准中对语言形式与策略能力（如任务回应、连贯性）做了精细拆解，并明确将学术写作作为核心构念；而 HSK 3.0 在这些维度的具体操作化落实上尚需进一步厘清。

这种国际化趋同与局部设计的差异，恰恰凸显了对 HSK 3.0 开展系统性效度验证的迫切性。明确 HSK 3.0 书写部分究竟考查了哪些潜在能力，并判定其任务设计是否真正实现了从“形式技能”向“交际策略”的有效跨越，仅凭单版分析难以得出科学结论，必须将其置于新旧对比的动态视野中。基于此，本研究重点探讨以下两个核心问题：（1）设计构念视角下，HSK 2.0 和 HSK 3.0 考试的书写部分在任务特征与潜在构念覆盖上有何异同？（2）HSK 3.0 书写部分的构念演变对国际中文书写教学及未来测试优化带来哪些启示？

文献综述

在语言测试学领域，构念效度理论与二语写作测评框架的演进为本研究提供了核心理论坐标。

效度理论的演进与测试辩证

构念效度是语言测试的核心。效度理论从早期 Gulliksen (1950) 的经典测量理论，发展至 Messick (1989, 1995) 的统一效度观，确立了构念效度在测试解释中的统领地位，并强调从“内容”与“实质”层面去验证测试任务与目标构念的一致性。近年来，Kane (2013) 进一步提出了“基于论证的效度验证模式 (Argument-based Approach)”，强调效度并非测试本身的固有属性，而是对测试分数解释和使用的逻辑论证。这一前沿视角的演变提示我们，对 HSK3.0 这样的大规模新兴测试，不能仅停留在题型特征的表面描述，而必须深入解构其任务要求背后的“潜在构念”，并通过多维视角的三角互证来审视其效度逻辑。

大规模二语写作测试改革与构念演变

纵观国际语言测试宏观图景，二语写作测试的构念经历了从“结构主义”向“交际与学术功能”的深刻转型。Weigle (2002) 指出，早期的写作测评往往被窄化为对语法和词汇准确性的形式测量；而近年来，以雅思、托福为代表的主流测试纷纷引入数据概括、文献综合等复杂任务，强调在真实学术语境中的交际功能与语用得体性 (Knoch & Macqueen, 2020)。HSK3.0 书写部分的改革正是顺应了这一国际化演变趋势。

二语书面产出与任务复杂性研究

在二语习得与测评实证领域，学习者的书面产出表现深受测试任务特征的制约。任务型语言教学 (TBLT) 的经典理论 (Skehan, 1998; Ellis, 2003) 表明，真实交际情景的介入与意义优先的任务特征能够有效激发学习者的高阶语言策略；而 Robinson (2001) 在其“认知假说”中提出的“三元组成框架”进一步指出，任务的内在认知复杂度 (如认知负荷与情景复杂度) 与任务条件 (如语言资源需求) 是影响学习者语言产出的关键因素。目前，学界对全球汉语学习者的书写能力发展状况已有初步探讨 (丁安琪等, 2022)。鉴于 HSK3.0 于 2026 年初才正式进入试点开考阶段，相关样题发布不久，目前学界关于 HSK2.0 与 HSK3.0 书写部分演变的系统对比研究尚处于起步阶段，更鲜有研究将上述二语习得与任务设计的经典理论引入对新版考试改革的深度剖析中。此外，诸如“以证据为中心的设计 (ECD)”等现代测评建模理论，以及“布鲁姆认知分类学”等教育心理学框架，为精细解构测试任务提供了科学的量化标尺，但在目前的 HSK 研究中同样缺乏应用。基于此，本研究试图打通测试学理论、认知科学与二语习得框架，对 HSK 新旧版本的书写构念展开系统对话，以期为 HSK3.0 的全面推行与教学适应提供具有前瞻性的学理支撑。

HSK2.0 和 HSK3.0 书写部分题型及题量对比分析

HSK2.0 与 HSK3.0 书写部分在题型配置、主客观题量分布及能力考查的难度梯度上均呈现出显著的结构性演变。为更直观地呈现这些变化，本研究将两版考试各等级的具体题型、主客观属性及总字数要求进行了系统梳理 (详见下页表 1)。

表 1. HSK2.0 与 HSK3.0 各等级书写部分题型、题量及字数要求对比

等级	HSK2.0 书写部分 题型构成与题量	总字数 要求	HSK3.0 书写部分 题型构成与题量	总字数 要求
一级	无	/	无	/
二级	无	/	[客] 选择汉字部件（5 题） [客] 写汉字（5 题）	/
三级	[客] 完成句子（5 题） [客] 写汉字（5 题）	/	[客] 写汉字（5 题） [主] 看图词写句子（5 题）	/
四级	[客] 完成句子（10 题） [主] 看图词写句子（5 题）	/	[主] 看图词写句子（5 题） [主] 根据话题写短文（1≥80 字 题）	
五级	[客] 完成句子（8 题） [主] 给定词写短文（1 题）≥160 字 [主] 看图写短文（1 题）		[主] 多图看图写作（1 题， 100 字） [主] 文字要求写短文（1≥300 字 题，200 字）	
六级	[主] 缩写（1 题）	≥400 字	[主] 应用文写作（1 题，150 字） [主] 文字要求写短文（1≥450 字 题，300 字）	
七 - 九 级	（无此等级）	/	[主] 图表写作（1 题，200 字） [主] 限时话题作文（1 题，≥800 字 600 字）	

注：表中[客]代表“客观题”，[主]代表“主观题”。

结合表 1 的统计，两版考试的演变特征可归纳为以下三个维度：

第一，题型更迭体现了测查重心的转移。HSK3.0 淘汰了注重局部语法的“完成句子”和机械记忆负荷过重的“缩写”题型。一方面，新增“选择正确的汉字部件”题，将考查起点下沉，细化了基础书写能力的颗粒度；另一方面，中高等级增加了“应用文写作”与“图表写作”等实用文体，标志着测查重心由传统的“结构主义”形式结构，向真实语境下的交际功能产出转移。

第二，从题目数量及题型性质来看，客观题呈前置与削减趋势，主观题比重显著攀升。在 HSK2.0 中，客观题贯穿三至五级；而在 HSK3.0 中，客观题的出现时间由三级前置至二级，且从五级开始彻底退出测查框架。与此同时，主观题（如看图写句、写短文）的引入节点均不同程度地提前（如写短文由 HSK2.0 的五级前置至 HSK3.0 的四级）。这表明 HSK3.0 更早、更强地介入了对学习者自主输出能力的测查。

第三，难度梯度设计更为精细，整体产出要求大幅提升。纵向来看，两版考试均遵循从字词识别、单句表达到语篇建构的递进规律，但梯度设置有所不同。HSK2.0 在五级（总计 160 字）至六级（缩写 400 字）之间存在一个高达 150%的字数跨度。HSK3.0 则通过增设级别，将难度分散为三个增长梯度（四至五级、五至六级、六至七-九级）。横向比较而言，HSK3.0 的整体输出要求更为严苛。需要客观指出的是，虽然 HSK3.0 试图平缓难度阶梯，但从四级（单任务，80 字）到五级（双任务，总计 300 字），在任务数量与总产出字数双重叠加下，其增幅仍高达 275%，这一局部跨度对中

级学习者而言无疑构成了较大的产出压力。总体而言，HSK3.0 实现了难度颗粒度的精细化与测查天花板的拔高，对中高等级学习者的语言资源调动与篇章产出能力提出了更真实的挑战。

基于构念效度的 HSK2.0 和 HSK3.0 书写部分对比分析

根据经典测试理论（CTT），在教育测量学中，信度和效度是两大核心质量指标。信度主要是控制随机误差，保证测量结果的稳定性、一致性、可重复性；效度则是指是否真正测到了想要测的目标（Gulliksen, 1950）。HSK 作为国家标准化的汉语水平考试，各等级的考试试题主客观题配比适度，客观题具有答案唯一性，主观题的评判也有一定的标准，整体具有较高信度。但作为一项语言测试，是否测到了想要测的目标，即构念效度问题，值得进一步思考。

本部分基于 Messick（1989, 1995）的构念效度统一框架与 Bachman（1990）的交际语言能力模型（CLA），从构念的代表性与覆盖面、任务特征与构念的内在关联性、难度梯度与构念发展层级的适配性三个维度对 HSK2.0 与 HSK3.0 书写部分进行对比分析。三个维度之间存在逻辑递进关系。构念的代表性与覆盖面是效度的基础条件，决定测试是否全面覆盖目标构念的子维度；任务特征与构念的内在关联性是效度的深层验证机制，反映题型能否在外显的情境功能与内隐的认知加工两个层面真实触发目标构念；难度梯度与构念发展层级的适配性则是效度的结构验证，检验测试的内部梯度是否与构念水平的发展顺序相一致。三者共同构成构念效度的系统分析框架。

书面语言交际能力构念的维度界定与理论推演

构念效度是语言测试的核心与基石。根据 Messick（1989, 1995）更新的构念效度统一框架，效度验证包含内容（content）、实质（substantive）、结构（structural）、泛化（generalizability）、外部（external）与结果（consequential）六个层面。就 HSK 考试书写部分而言，本研究主要聚焦于探讨其内容层面（即测试题型是否全面、代表性地覆盖了目标构念的边界）与实质层面（即测试任务特征能否有效触发目标构念预期的内部认知与交际过程），并在反思与展望阶段，向外延伸至测试的结果层面（教学反拨效应）、结构层面（评分标准适配）与外部/泛化层面（国际标准参照与学术情境推广）。因此，明确测试的目标构念——“汉语书面语言交际能力”的具体内涵，是开展效度分析的前提。

为科学界定该构念，本研究引入 Bachman（1990）的交际语言能力模型（CLA）。该模型指出，交际语言能力主要由语言能力、策略能力和心理生理机制构成。其中，语言能力分为组织能力（含语法能力、语篇能力）和语用能力（含施为能力、社会语言能力）；策略能力（即认知与元认知策略）则是语言知识转化为实际交际表现的执行“桥梁”；心理生理机制则指语言产出的神经与生理载体。

为使这一宏观理论模型适应汉语书写测试的特殊性，本研究对其进行了学科化与情境化的“拆解”与“合并”，推导出以下五个书面语言交际子构念：

第一，书面字词应用构念与书面句法应用构念。这两个子构念分化自 Bachman 模型中的“语法能力”。鉴于汉语是非形态语言，其汉字（形体结构）与语法（词序与虚词）分属两套独立平行发展的编码体系，认知路径存在本质差异（徐通锵，2008）；而留学生的书写困难则集中体现于从部件到整字的加工环节（张旺熹，1990），这与句法层面的学习难点源于不同的认知基础。因此，本研究将广义的语法能力拆解为：

- （1）书面字词应用构念（包含汉字规范书写、部件识别与组合、词语的正确选择）；
- （2）书面句法应用构念（包含句法结构的正确性、复杂性及句式运用的多样性）。

第二，书面语篇建构构念。该构念直接映射自 Bachman 模型中的“语篇能力”，主要关注超越句子层面的意义组织，包括语段衔接手段的运用、逻辑连贯以及篇章结构的完整性。

第三，书面交际语用构念。该构念直接映射自 Bachman 模型中的“语用能力”（主要包含施为能力与社会语言能力），并将其置于书面交际的真实语境中进行深化。在真实的汉语书面交际（如应用文、图表分析）中，语用规则的最终指向是交际效用的达成：学习者不仅需要通过施为能力实现特定的交际功能（如说明、论证、请求），还需通过社会语言能力选择恰当的语体、遵循相应的文化及格式规范。因此，本研究将其界定为书面交际语用构念，主要包含情景适配性、语体得体性及书面交际功能的实现。

第四，策略构念。该构念对应 Bachman 模型中的“策略能力”，被置于能力链条的高阶位置。随后，Bachman 和 Palmer (1996) 在测试实践中进一步将该能力操作化为一系列元认知执行成分。在写作任务中，它具体化为元认知策略（如任务前的计划、任务中的监控与调节、任务后的评估）和认知策略（如对信息的选择性注意、推理、概括与整合等）。此外，由于心理生理机制主要涉及书写动作等生理载体，不属于本研究语言认知与交际效度探讨的核心范畴，故在此不作独立构念提取。

上述五个子构念紧扣 CLA 模型，逻辑严密地构成了一条从低阶形式解码（字词、句法）到高阶意义建构（语篇、语用），再到宏观执行调控（策略）的汉语书面语言能力发展完整链条。下文将基于这一理论演化而来的构念框架，对 HSK2.0 和 HSK3.0 书写部分展开系统对比。

核心题型的多维构念效度对比

构念的代表性与覆盖面：从“窄化覆盖”到“全维度覆盖”

构念覆盖度是指测试题型所测量的子构念是否全面代表了理论目标构念的各个维度。为避免主观定性，本研究引入 Mislevy 等人（2003）的“以证据为中心的设计”（Evidence-Centered Design, 简称 ECD）理论作为构念覆盖强度的客观评判依据。

ECD 理论指出，测试任务所激活的知识、技能与能力（Knowledge, Skills, and Abilities, 简称 KSAs）并非平行分布，而是存在严密的测量层级，具体包含三个部分：其一为“核心目标（Focal KSAs）”，即测试意图测量的最核心、最显性的属性；其二为“支撑要求（Supporting KSAs）”，即完成测试任务必不可少的先决条件与基础技能；其三为“边缘/附带要求（Ancillary/Incidental KSAs）”，即任务执行过程中偶然触发、但不作为核心测评目标的能力。

严格依据上述 ECD 理论的构念分层模型，本研究通过系统分析官方考试大纲中的任务指令与预期表现，将题型对子构念的覆盖强度顺理成章地推导并界定为以下三个标准：

- A. 强覆盖（✓✓）（严格对应 Focal KSAs）：指该构念是任务要求的核心显性目标。若考生不具备该构念（如应用文写作中缺乏语体适配性、图表写作中缺乏认知归纳策略），将直接导致评估维度的交际任务失败或严重偏离交际目的。
- B. 覆盖（✓）（严格对应 Supporting KSAs）：指该构念是完成任务必不可少的隐性支撑条件。它虽非题目唯一考查核心，但考生必须调用该能力（如进行高阶语篇建构时必须使用合乎语法的句子与语篇衔接手段）才能产出合格语料。
- C. 弱覆盖/隐性覆盖（●）（严格对应 Ancillary/Incidental KSAs）：指该任务可能局部或偶然触发该构念，但不作系统性量化要求，且该构念的弱化或缺失不影响任务的基础完成（如写汉字时的基础自我监控，或缩写中极其微弱的交际意图）。

表 2 展示了基于上述 ECD 构念分层标准，HSK2.0 与 HSK3.0 各主要题型对五个子构念的系统覆盖情况。

表 2. HSK2.0 与 HSK3.0 书写部分题型对子构念的覆盖对比

版本	题型	书面 字词 应用	书面 句法 应用	书面 语篇 建构	书面 交际 语用	策略	
						认知	元认知
HSK 2.0	写汉字	✓✓				●	
	完成句子	✓	✓✓			●	
	根据图片和词语写句子	✓	✓✓			●	
	给定词语写短文（五级）	✓✓	✓	✓✓	●	✓	●
	单图写短文（五级）	✓	✓	✓✓	●	✓	●
	缩写（六级）	✓	✓	✓	●	✓	●
HSK 3.0	选择正确的汉字部件	✓✓				●	
	写汉字	✓✓				●	
	根据图片和词语写句子	✓	✓✓			●	
	话题短文（四至六级）	✓	✓	✓✓	✓	✓	✓
	多图看图写作（五级）	✓	✓	✓✓	●	✓✓	✓
	应用文写作（六级）	✓	✓	✓	✓✓	✓	✓✓
	图表写作（七-九级）	✓	✓	✓	✓✓	✓✓	✓✓
	限时话题作文（七-九级）	✓	✓	✓✓	✓✓	✓✓	✓✓

由表 2 可见，在 ECD 构念分层模型的审视下，HSK2.0 书写部分呈现明显的构念窄化特征。其绝大多数题型的核心测查目标（Focal KSAs，即强覆盖✓✓）被锁定在书面字词应用与书面句法应用等低阶形式技能上；对语篇建构构念的核心覆盖虽在五级有所触及，但其覆盖层级单一，且未能在此基础上向语用与策略等更高阶构念延伸；而书面交际语用构念则整体处于边缘化的附带状态（●）；认知与元认知策略也大多仅作为基础支撑或偶然触发存在。整体而言，HSK2.0 未能将高阶语言交际能力纳入核心测量体系，代表性不足。

相比之下，HSK3.0 实现了“部件→字词→句子→语篇→交际→策略”的全维度覆盖。一方面，新增部件题将考查起点下移，夯实了基础支撑（Supporting KSAs）；另一方面，通过应用文、图表写作等交际任务，首次将书面交际语用构念、高阶认知策略与元认知策略由过去的边缘状态，正式拔高为测试的核心目标（强覆盖✓✓）。总体而言，HSK3.0 的构念覆盖度显著优于 HSK2.0，彻底解决了前者交际构念窄化的问题。

任务特征与构念的内在关联性：情境真实性与认知加工的双重触发验证

ECD 分层模型虽能标定题型的构念覆盖边界，但覆盖不等于触发。一个题型即使被贴上“语用”或“策略”的强覆盖标签，若其任务情境脱离真实交际（只需形式操练即可完成），或其内在认知要求停留于机械记忆，则构念效度依然落空。因此，本部分引入 Weir（2005）的“社会-认知效度验证框架”（Socio-Cognitive Framework），从“外显的情境功能”与“内隐的认知加工”两个互补层面，验证各题型是否在任务执行层面真实触发了目标构念。前者确保任务具备真实的交际情境与意图，后者确保任务触发了与目标构念层级相匹配的认知加工过程。基于此，本研究分别从以下双重视角对测试任务进行解构。

首先，在外显的情境功能维度，本研究依据任务型语言教学（TBLT）对有效交际任务的核心界定（Skehan, 1998），将其操作化为两项具体的判断标准：第一，是否要求考生主动建构意义；第二，是否包含真实/拟真的社会交际情境。结合这两项标准来看，HSK2.0 书写题型大多停留在“浅层形式关联”。例如，“完成句子”主要要求考生进行词语的语法排序，属于脱离语境的形式操练，既不要求自主建构意义，也缺少真实的交际受众。高等级的“缩写”虽然触及语篇，但本质仍是对给定信息的记忆与复述，考生的自主表达空间极度受限。与之根本不同的是，HSK3.0 实现了向“深层功能关联”的转型。其高等级任务（如应用文写作、图表写作）均设置了明确的交际意图与特定的交际对象，强制要求考生根据情景规范（如礼貌原则、正式语体）来进行意义的建构与输出，这契合了“意义优先”与“真实世界关联”的标准。上述分析表明，在外显情境层面，HSK3.0 已具备触发语用与策略构念的交际生态前提。然而，任务是否在认知加工层面真实激活了相应构念，还需从内隐维度作进一步验证。为此，本研究引入安德森等人修订的“布鲁姆教育目标分类学”（Anderson & Krathwohl, 2001）作为测定依据，以客观丈量任务所激发策略的认知深度。该理论将认知过程严密划分为由低到高的六个层级：记忆（Remember）、理解（Understand）、应用（Apply）属于低阶认知技能（LOCS）；分析（Analyze）、评价（Evaluate）、创造（Create）则属于高阶认知技能（HOCS）。本研究以此为理论标尺，对两版考试主要题型的内在“隐蔽认知加工链条”进行了逆向拆解，并实现了核心策略与布鲁姆认知层级的严格映射（见下页表 3），为触发效度提供认知层面的直接证据。

结合表 3 可以清晰发现，在权威认知框架的映射下，HSK2.0 题型的认知加工链条绝大多数被锁定在“记忆、理解、应用”这一低阶认知技能（LOCS）区间，考生主要需完成外在语言形式的提取与匹配。

与之形成鲜明对比的是，HSK3.0，特别是其中高等级的交际任务，实现了向高阶认知技能（HOCS）的整体跃迁。以七至九级的“图表写作”为例，考生必须经历极为复杂的认知嵌套流程：识读数据差异（对应布鲁姆“分析”层级）、论证核心观点并预判交际效果（对应“评价”层级）、最终完成具有严谨逻辑与语体规范的篇章构建（对应“创造”层级）。可见，HSK3.0 的任务设计并非表面文体的更迭，而是通过精细的内在任务链设计，迫使考生突破低阶形式技能的局限，最终实现了测试任务在“外显交际功能”与“内隐认知深度”上的高度统一。

需说明的是，本小节聚焦于任务情境的交际真实性与认知加工的层级深度，旨在验证构念是否被真实触发；而第三小节对情景复杂度的考察，则侧重于该复杂度在不同等级间的梯度分布是否契合构念发展顺序——二者视角有别，前者为“是否触发”的质性判断，后者为“如何递进”的结构考察。

难度梯度与构念发展层级的适配性：从“技能难度”到“能力难度”

难度梯度设计是否合理，取决于其是否与“字词→句法→语篇→语用→策略”的真实能力发展顺序相契合。为克服指标分类的主观性局限，本研究引入 Robinson（2001）的认知假说（Cognition Hypothesis）及其“三元组成框架”。本研究将其严格映射为测试任务的三个评价维度：认知负荷、语言资源需求与情景复杂度。

为确保评判的客观性，下页表 4 中的“高、中、低”层级并非依赖个人主观感知，而是通过理论框架对《等级标准》的量化指标及官方任务指令进行严格的“操作化界定（Operationalization）”，具体判定依据如下：

表 3. 主要书写题型的认知加工链条与布鲁姆层级映射

版本	题型	内在认知加工链条	调用的核心策略 (认知/元认知)	布鲁姆认知 层级界定
共有	写汉字	提示理解 → 字形检索 → 书写执行	提取、回忆 / 无	记忆、应用 (低阶)
	根据图词 写句子	图义提取 → 词语整合 → 句法构建	观察、匹配 / 无	理解、应用 (低阶)
HSK 2.0	完成句子	词语排序 → 句法套用 → 通顺核查	匹配、排序 / 无	应用 (低阶)
	给定词语 写短文	词汇释义 → 内容拼凑 → 句子展开	联想、组织 / 基础评估	理解、应用 (低阶)
	单图写短文	画面信息提取 → 情景 推断 → 叙事	观察、推理 / 基础评估	理解、分析 (低阶跨中阶)
	缩写 (六级)	原文理解 → 主干筛选 → 重组转述	概括、转述 / 记忆监控	理解、应用 (低阶)
	选择汉字 部件	部件识别 → 位置判断 → 干扰排除	识别、区分、 匹配/无	记忆、理解 (低阶)
HSK 3.0	话题短文 (四级)	话题理解 → 素材提取 → 框架搭建	联想、组织 / 任务前计划	应用、创造 (低阶跨高阶)
	多图看图 写作	多图排序 → 信息合并 → 连贯叙述	排序、归纳 / 逻辑监控	分析、创造 (高阶)
	应用文写作	交际分析 → 内容组织 → 语体选择	提取、组织 / 读者意识调控	分析、创造 (高阶)
	图表写作	数据比对 → 趋势分析 → 观点论证	分析、推理 / 计划、交际调 控	分析、评价、创造 (高阶)
	限时话题 作文	快速立意 → 提纲规划 → 限时产出	推理、发散 / 进度评估	分析、评价、创造 (高阶)

表注：策略调用中“/”前为认知策略，“/”后为元认知策略；“低阶/高阶”的严密划分依据为 Anderson & Krathwohl (2001) 的布鲁姆认知模型。布鲁姆层级的判定依据为任务的信息输入类型与自主建构空间：（1）“单图写短文”的画面提供了完整的事件要素，考生主要在观察、理解与转述层面操作，其终点为“分析（中阶）”；（2）“话题短文”仅提供抽象话题，无具体内容支撑，考生需从零生成意义结构，故触及“创造（高阶）”；（3）“多图看图写作”需自主判断图间逻辑顺序并将离散画面整合为连贯叙事，涉及跨图逻辑推理与意义建构，故亦归为“创造（高阶）”。

A. 认知负荷：以完成目标任务所必需的“内在认知加工步骤”为推演依据。尽管官方考试指令仅呈现外在的行为要求（如“根据图表写一篇短文”），但本研究依据认知加工理论，对成功达成该交际目的所隐蔽包含的思维链条进行了客观拆解。表中提取的认知特征反映了清晰的进阶轨迹：从基于给定情景的单向信息输出（如四级的“单一主题叙述”），递进至需要跨媒介或跨要素整合的逻辑推演（如五级的“多图逻辑整合”），直至高阶的多重推理嵌套任务（如七至九级的“数据解读+观点论证”）。

- B.语言资源需求：依据《等级标准》附录 A “语法等级大纲”与附录 C “词汇等级大纲”。例如：四级需求对标大纲中的“基础词汇与基本复句”；五级需求对应大纲中“体标记”在复杂语境中的运用、篇章“衔接手段”、“时间关联词”及复杂句式；六级及以上则对应大纲中要求掌握的“书面语体标记”、“礼貌策略表达”及复杂句群。
- C.情景复杂度：依据《等级标准》附录 B “话题任务内容”。“低”为无特定交际受众的通用个人话题（对应《等级标准》四级日常话题）；“中”为附带一定情景条件的生活叙事；“高”为具有明确社会交际对象限制、文体格式规范及跨文化语体要求的真实应用场景（完全匹配《等级标准》六级及以上的复杂社会与专业话题要求）。

基于上述透明、可回溯的推演标准，本研究对两版考试的难度属性进行了理论解析。在 Robinson 的框架审视下，HSK2.0 的书写任务（如词语排序）因严重脱离真实语境，其情景复杂度与认知负荷均处于极低水平，难度区分主要依赖孤立的词汇量与语法点叠加，属于典型的“技能难度”。相比之下，HSK3.0 中高等级任务则真正实现了向“能力难度”的跃迁，其具体难度属性解析详见表 4。

表 4 HSK3.0 书写部分五级至七-九级题型的认知负荷、语言资源需求及测试构念

等级	题型	认知负荷	语言资源需求	情景复杂度	主要构念
四级	话题短文（80 字）	单一主题叙述	基础词汇与基本复句	低	语篇建构
五级	多图写作+话题作文（≥300 字）	多图逻辑整合+观点论证	衔接手段、体标记与时间关联词、复杂句式	中	语篇+策略
六级	应用文+话题作文（≥450 字）	交际意图+格式规范+观点论证	书面语体标记、礼貌策略、复杂句群	高	语用+元认知
七-九级	图表解读+话题作文（≥800 字）	数据解读+观点论证	图表阐释、抽象词汇、复杂句群、高密度信息组织	高	语用+策略

注：难度指标的理论推演依据 Robinson（2001）任务复杂性框架；“语言资源需求”与“情景复杂度”已完成与 2021 版《等级标准》附录相关大纲的特征映射。

由表 4 可见，HSK3.0 实现了“能力难度”的递进升级。五级逻辑整合增强，认知策略显著参与；六级增加交际意图与语体规范；七至九级叠加数据解读与限时规划，认知负荷和情景复杂度均处于高位，与理论上的构念发展层级一一对应。正如 Swain（1993）的“输出假说”所言，这种更高认知要求与更长篇幅的书面输出，能进一步逼迫二语学习者进行深层句法加工，从而促进语言习得。但正如前文实证数据所揭示的，HSK3.0 四级（单任务 80 字）至五级（双任务共 300 字）的跨度因任务数量增加与

逻辑整合要求叠加，可能导致认知负荷跳跃式增长，这是未来测试优化与教学适应需重点关注的方向。

基于专家评判与学习者感知的实证补充

研究对象与方法

为进一步验证前文基于文本分析的理论推导，本研究采用混合研究方法，从“专家评判”与“学习者感知”两个维度进行初步的三角互证。在样本选择依据方面：在专家评判环节，本研究采用目的性抽样（**Purposive Sampling**），严格选取了 3 名具有 10 年以上一线国际中文教学经验、且长期从事 HSK 写作辅导的资深教师（以下分别简称教师 A、B、C）开展半结构化访谈。虽然专家数量有限，但其深厚的教学背景能够为新题型任务特征提供深度的“专家效度（**Expert Validity**）”支持。在学习者感知环节，受限于 HSK 3.0 为 2026 年 1 月刚进入试点的新兴测试，大规模曾参加过实测的受试群体尚未形成。因此，本研究采用方便抽样（**Convenience Sampling**），选取了 23 名正在学习 HSK 4 级的留学生。该群体国籍涵盖土库曼斯坦、坦桑尼亚、也门等 8 个国家，具有较好的生源代表性；更为重要的是，中级学习者正是未来面临向高阶难度（如 HSK 5 级双任务）跨越的直接利益相关群体，他们对难度梯度的感知具有极强的典型性。在数据处理与局限性方面：需要明确指出，由于受试者普遍未系统完成过 HSK 3.0 的样题，问卷结果仅代表基于任务说明的“主观感知难度”。本研究中的问卷数据仅作为描述性统计（**Descriptive Statistics**），专家访谈也仅提炼核心主题。小规模实证数据主要发挥辅助论证与三角互证的功能，用于印证理论推导的合理性，其结论不作绝对化推广。

专家访谈结果分析

针对访谈语料，本研究采用主题分析法进行提炼。总体而言，三位专家对 HSK3.0 书写部分的改革方向持高度肯定态度，但在局部设计上存在一定分歧，具体体现在以下三个主题：第一，构念覆盖更广，但学术维度仍显不足。三位教师均认同 HSK3.0 的任务设计更贴近真实交际，弥补了 HSK2.0 的短板。大家一致认为“图表写作”能有效诱发数据解读等认知策略，而“应用文写作”能真实测查语用得体的性。然而，教师 B 特别指出，HSK3.0 的任务情境仍偏向通用汉语，针对来华留学生的“学术写作”能力构念尚未被独立触及。第二，难度梯度总体递进，但局部跨度存在争议。针对 HSK3.0 四级至五级写作任务字数从 80 字跃升至 300 字的设置，教师间看法不一。教师 A 与教师 C 认为这一跨度“过大，缺乏过渡”，易造成中级学习者认知负荷陡增；教师 B 则基于更高的产出期望，认为“跨度不够，四、五级均应有更高要求”。这表明，基于理论推演的难度阶梯在实际教学体感中仍需进一步调试。第三，肯定交际导向，需警惕应试化反拨。在反拨效应方面，受访专家一致认同 HSK3.0 的设计理念已向雅思等国际标准看齐，但教师 A 与教师 C 担忧，过于聚焦“图表整合”等特定任务可能在实际教学中催生新式“套句模板”的应试训练。不过教师 B 认为，适度的套用在真实职场中确有应用价值。此外，专家们一致强调，主观题比重的大幅增加，要求考试方在评分标准的透明度与阅卷员培训上积极借鉴国际经验。

学习者难度感知调查

针对 23 名 HSK4 级学习者的问卷调查结果表明，各类书写题型的平均感知难度（1-5 分制）排序，与前文基于布鲁姆教育目标分类学推导的“构念层级”呈现出高度一致性。整体感知难度可明显划分为三个梯队：

首先是低难度梯队（均分<2.5），依次为连词成句（1.7）、写汉字（1.96）、选择汉字部件（2.17）与图词写句子（2.3）。这印证了前文的分析，即此类任务主要调用“记忆、理解”等低阶认知技能（LOCS），学习者认知负荷较小。这从学习者感知层面印证了ECD判定中低阶题型的构念覆盖与认知负荷的一致性。

其次是中等难度梯队（均分 2.8-3.9），主要集中在语篇建构与基础交际任务上，如给定词语写短文（2.8）、四级话题短文（3.22）、多图看图写作（3.57）及六级应用文与话题作文（均为 3.83）。随着任务对信息整合与语体规范的要求提升，难度感知呈平缓上升趋势。在实用性方面，“应用文写作”被多数受试者认为最能切实提高真实汉语写作能力，这侧面印证了该新题型在测查语用构念上的高保真度。学习者对应用文写作实用性的认可，从使用者视角佐证了该题型在语用构念测查上的高生态效度。

最后是高难度梯队（均分>4.0），包含限时话题作文（4.17）、缩写（4.22）与图表写作（4.22）。受试者普遍认为 HSK3.0 新增的“图表写作”极具挑战，原因在于其“需要逻辑思维和分析能力同时进行”，这完全契合布鲁姆教育目标分类学中“分析、评价”等高阶认知技能（HOCS）的复杂嵌套。值得关注的是，属于 HSK2.0 的“缩写”题型，其所测构念层级相对较低，但因限时阅读带来的“工作记忆负荷”过大，被学习者感知为最难并列第一。这侧面提示我们：测试的外在感知难度不仅来源于目标构念本身的层级，还深受任务实施条件（如时间限制、记忆负荷）的强烈干预。

结论与启示

本研究基于 Messick（1989, 1995）构念效度理论与 Bachman（1990）交际语言能力模型，并引入 ECD 构念分层模型与布鲁姆认知分类学作为操作化标尺，对 HSK2.0 与 HSK3.0 书写部分围绕五个子构念展开系统对比，并辅以专家访谈与学习者问卷进行初步三角互证。

主要研究结论

第一，在构念代表性与覆盖面维度，HSK2.0 呈现明显的窄化特征，其核心测查目标（Focal KSAs）局限于字词与句法；HSK3.0 则实现了从汉字部件到语篇、交际语用及策略构念的全维度覆盖，有效拓展了测评边界。第二，在任务特征与构念的内在关联性维度，HSK2.0 多停留在浅层形式关联，加工过程被锁定在低阶认知技能（LOCS）区间；HSK3.0 则转向深层功能关联，通过真实的交际情景设计，迫使考生调用推理、归纳与交际调控等策略，成功实现了向高阶认知技能（HOCS）的整体跃迁。第三，在难度梯度与构念发展层级的适配性维度，HSK3.0 的整体梯度设计更为精细。其任务的认知负荷、情景复杂度与语言资源需求，从“技能难度”向“能力难度”升级，与理论上的构念发展层级高度适配。但需指出，四级（80 字）至五级（双任务 300 字）的局部跨度因任务叠加导致增幅达 275%，存在认知超载风险，提示梯度设计的精细化仍有优化空间。

教学实践与未来测试优化启示

HSK3.0 的系统性升级充分体现了能力本位与交际本位的评价导向。然而，初步的实证数据在为理论推演提供支撑的同时，也揭示了测试在未来实操中面临的复杂协变量与挑战。基于此，本研究提出以下三方面启示：

第一，立足全维度构念，警惕应试化教学的负面反拨。HSK3.0 将测量范围拓展至高阶交际与策略构念，提示教学必须突破“重字词、轻语用”的传统局限。教师需高

度警惕新题型可能催生的“新式应试化”现象，如不应将“图表写作”简化为固定句型的机械套用，而应培养学习者真实的意义建构能力，让书写回归交际本质。

第二，遵循能力递进规律，平缓局部认知跨度。实证揭示，四级（单任务 80 字）至五级（双任务共 300 字）因任务叠加导致的总字数陡增，易致中级学习者面临“认知超载”。教学中应自主增加过渡性训练以合理控制篇幅增长节奏；同时建议 HSK3.0 未来修订时，适当调整五级起点的字数要求或增设平滑的过渡任务，以更精准地匹配学习者生理与心理发展的真实节奏。

第三，拓展测评构念边界，填补学术缺位并细化评分标准。当前 HSK3.0 在独立学术写作（如文献观点提炼、引用转述）维度的测查依然缺位。建议未来在高级别之上开发独立的“HSK 学术（Academic）”模块，以提升其在真实大学学术情境中的国际预测力。此外，针对高阶构念（如语用得体的性）在实测中面临的计分信度存疑及“评价窄化”等严峻挑战，未来需重点关注评分标准的可操作化细化。研究者亦应结合阅卷实测数据与有声思维协议，深入探究高阶构念在实际评分中的真实落实情况。

参考文献

- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Ding, Anqi (丁安琪), Li, Peize (李佩泽), Yu, Yan (于艳), & Cheng, Haiting (程海婷). (2022). 全球汉语学习者汉语能力表现的现状与问题——基于 2018 年汉语水平考试(HSK)大数据的分析 [Current Situation and Problems of Global Chinese Learners' Chinese Proficiency Performance: An Analysis Based on Big Data of the 2018 HSK]. *对外汉语研究* [Studies in Chinese as a Foreign Language], (1), 1-11.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.
- Gulliksen, H. (1950). *Theory of mental tests*. Wiley.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Knoch, U., & Macqueen, S. (2020). *Assessing English for professional purposes*. Routledge.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Ministry of Education & National Language Commission (教育部、国家语言文字工作委员会). (2021). *国际中文教育中文水平等级标准 (GF 0025-2021)* [Chinese Proficiency Grading Standards for International Chinese Language Education]. 北京语言大学出版社 [Beijing Language and Culture University Press].
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27-57.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
- Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *Canadian Modern Language Review*, 50(1), 158-164.

- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- Xu, Tongqiang (徐通锵). (2008). 汉语字本位语法导论 [An Introduction to Chinese Character-based Grammar]. 山东教育出版社[Shandong Education Press].
- Zhang, Wangxi (张旺熹). (1990). 从汉字部件到汉字结构——谈对外汉字教学 [From Chinese Character Components to Character Structure: On Teaching Chinese Characters as a Foreign Language]. *世界汉语教学* [Chinese Teaching in the World], (2), 112-120.

李蓉，湖北荆州人，江汉大学人文学院讲师，硕士，研究方向为国际中文教育。