Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2020, 2(4), 175-183
Published Online November 2020 in Hans. http://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2020.24017

Hans汉斯

眼动技术在语言测试研究中的应用现状和展望

许皖栋[1]，辜向东[2]

[1]香港理工大学，香港
[2]重庆大学，重庆
Email: xdgu@cqu.edu.cn, Wandong.xu@connect.polyu.hk

收稿日期：2020年11月3日；录用日期：2020年11月16日；发布日期：2020年11月30日

摘　要

眼动技术能够真实地记录语言加工活动中即时的眼球运动，将眼动技术应用于语言测试研究，有助于以实验心理学的研究方法为语言测试认知过程提供客观可靠的数据。本文简要介绍了近年国外语言测试研究中眼动技术的应用，相关研究主要关注：考生在测试环境和真实语境中认知过程的相似性，得分考生和失分考生认知过程的差异，特定题型所引发的认知过程，任务模态对考生认知过程的影响以及评分员的认知过程。在此基础上作者从多角度分析了相关研究存在的不足，以期为国内学者从事相关研究提供一定参考。

关键词

眼动技术，认知过程，语言测试研究，应用现状，展望

Applying Eye-Tracking Technology into Language Assessment Research: Status Quo and Prospects

Wandong Xu[1], Xiangdong Gu[2]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]Chongqing University, Chongqing
Email: xdgu@cqu.edu.cn, Wandong.xu@connect.polyu.hk

Received: Nov. 3rd, 2020; accepted: Nov. 16th, 2020; published: Nov. 30th, 2020

Abstract

Eye-tracking technology allows for authentic recordings of participants' online eye movements. Applying eye-tracking technology into language assessment research will help to provide an ob-

文章引用: 许皖栋, 辜向东. 眼动技术在语言测试研究中的应用现状和展望[J]. 国外英语考试教学与研究, 2020, 2(4): 175-183. DOI: 10.12677/oetpr.2020.24017

jective and reliable data source for studies on cognitive processes by means of experimental psychology. This paper offers a brief introduction of the status quo of language assessment research with the use of eye-tracking which mainly focuses on: the similarity of test takers' cognitive processes engaged in language testing context and real-life domain, the differences on cognitive processes of successful and unsuccessful test takers, cognitive processes elicited by particular test format, the influence of test modality on test takers' cognitive processes and raters' cognitive processes. In light of these studies, this paper pinpoints existing limitations from multiple perspectives and intends to provide valuable implications for researchers who are interested in related studies.

Keywords

Eye-Tracking, Cognitive Processes, Language Assessment Research, Status Quo, Prospects

1. 引言

自上世纪 80 年代以来，随着语言测试研究对过程证据的重视，考生的认知过程进入研究视野。基于考生的认知过程，语言测试研究者可以验证考试任务在多大程度上引发了考生在心理层面与预设构念相关的加工，深化人们对考试构念的认识，为考试效度提供依据[1]。考生如何作答的过程证据有助于支持分析考生的测试表现以及考试分数的解释与使用[2]。然而考生的认知过程难以直接观察，如何准确真实地记录和分析考生的认知过程一直是相关研究的难点和关键。

眼动技术的发展为语言测试，尤其是阅读测试的认知过程研究提供了新视窗。眼动仪可以实时记录考生在考试过程中的眼球运动，数据客观丰富且易于获得，其研究结果具有较高的生态效度[3] [4]。与以往认知过程研究中所使用的数据收集方法如有声思维、内省法、访谈和问卷调查相比，眼动技术在这方面有其独特优势，在语言测试领域的应用越来越多[5]。本文拟综述近年眼动技术在国外语言测试研究中的应用，以期为我国语言测试认知过程研究提供一定参考。

2. 眼动技术及其在语言测试研究中的应用原理

2.1. 眼动技术和眼动研究发展

眼动技术指通过眼动仪记录和测量被试在处理特定视觉信息时的眼动轨迹[6]。研究者可以利用眼动仪决定被试加工何种视觉信息材料，以及材料的呈现时间和呈现顺序[7]。通过眼动数据所反映的视觉信息选择模式，研究者可以对刺激材料本身展开探索，并对被试的认知加工过程进行有效推测[8] [9] [10]。基于相关实证研究的梳理分析，Rayner [11]将眼动技术在阅读以及其他信息加工任务中的应用分为三个主要时期：第一代眼动研究(1879~1920)源于 Javal 对眼球运动在阅读中作用的初始观察，主要揭示了基本眼球运动的事实性信息，包括眼跳潜伏期，眼跳抑制和知觉广度；第二代眼动研究(1930~1958)受实验心理学行为主义运动的影响，注重应用研究，主要关注刺激材料本身的特征，很少有研究者利用眼球运动探索认知过程；第三代眼动研究(1970~1998)的标志在于眼动技术本身的发展，眼动数据更加精确且更易于获得。目前，随着交互应用的出现，眼动研究迈入了第四代[12]，眼动设备可以呈现更为复杂的视觉刺激，在研究中的应用范围越来越广泛。

2.2. 眼动技术在语言测试研究中的应用原理

心理学家认为眼球运动与大脑加工联系密切，可以反映我们人类多种认知活动，如注意，阅读，预测，推理和记忆[13]。Just & Carpenter [14]提出了“眼脑一致”假说，认为眼球正在注视的内容即为大脑正在加工的内容，视觉过程和大脑认知加工过程之间没有延迟。大量的实证研究发现也基本证实了这一假说，奠定了眼动技术在研究中应用的理论基础。在语言测试研究中，通过眼动仪记录和分析考生完成测试任务时的眼球运动和眼动指标，可以准确地获知考生正在接收和加工的信息，从而推测考生即时的思维过程。

3. 眼动技术在语言测试研究中的应用

眼动技术记录和分析被试的眼动轨迹，极大地促进了对潜在认知过程的探究。近年来，语言测试学者逐渐认识到这一技术的优势，开始结合眼动技术和认知过程研究开展创新性尝试，从不同角度揭示语言测试情境下的认知过程。下文将从五个方面对眼动技术在语言测试领域中的应用研究进行综述。

3.1. 考生在测试环境与真实语境中认知过程的相似性

语言测试研究中的效度是一个整体概念，需要多方面的证据予以支持[15]。长久以来，效度研究都围绕着考试分数与所测构念之间的关系展开量化分析，忽视了考生的考试过程和语言能力构念本身的认知成分[16]。Weir [17]从考生的视角出发，提出了认知效度概念，将其作为社会认知效度验证框架的一个重要方面，旨在探究考生在考试环境中所经历的认知过程在多大程度上与真实语境中经历的认知过程相似。

剑桥高级英语证书考试(Cambridge English: Advanced, CAE)是一项全球范围内认可的学术英语能力测试，通过该考试的考生有机会参加英语国家院校的大学课程学习和学术研修，因此考生在 CAE 考试中经历的认知加工过程理应与真实学术环境下的认知加工过程相似。为了验证CAE的认知效度，Bax & Weir [18]结合眼动技术和问卷调查探究了考生在 CAE 阅读考试中所经历的认知过程。基于眼动数据的可视化分析和量化统计，Bax & Weir 识别和比较了被试在每道题项上的作答行为，包括对文本、题干和选项的视觉加工情况。结合问卷调查数据，Bax & Weir 对考生的认知加工过程做出了有效推测，发现 CAE 阅读考试任务成功地引发了考生在真实学术环境中一系列相似的认知过程，包括较高层次(如段落和篇章层次)上的信息加工。该研究结果为 CAE 阅读测试的认知效度提供了过程证据，一定程度上证明了眼动技术在语言测试效度验证研究中应用的可行性，弥补了以往质性研究方法存在的局限性。

3.2. 得分考生和失分考生认知过程差异

有研究表明，考生的考试表现与认知过程有着密切的联系，考试表现不同的考生在认知过程上存在一定的差异[19] [20] [21] [22] [23]。为了探究考试表现不同的考生在认知过程上的差异，Bax [24]结合眼动技术和访谈，分析比较了得分考生和失分考生完成雅思阅读测试的认知过程。基于得分考生和失分考生在回答每个题项时的注视时间和注视次数，Bax 发现得分考生和失分考生在多个题项上的注视指标存在差异。与失分考生相比，得分考生在题项对应的原文兴趣区上注视次数更多，注视时间更长。这说明得分考生能够更好地利用快速阅读的策略，迅速定位对应的原文并展开仔细阅读，获取有效信息。此外，该研究利用可视化数据，如注视轨迹图(Gazeplot)和热点图(Heatmap)，直观呈现了被试在答题过程中的注意分配情况，并结合访谈数据对考生当时的认知加工做出准确判断。结果表明雅思阅读测试题项能够有效区分得分考生和失分考生，不过该研究的比较分析仅限于较低层次的阅读认知加工，如词汇层面的匹配和句法层面的歧义处理。

3.3. 特定题型所引发的认知过程

完形填空是语言测试中一种常见的考试题型，要求考生联系上下文补全某一语篇中被删除的词汇。完形填空题型考查了何种语言能力构念？语言测试学界对这一问题并没有达成一致意见。一些研究者认为完形填空能够考查考生较高层次的语言加工能力[25] [26]，而另有研究者认为完形填空受语句范围的制约，只能考查较低层次语言技能，如语法和词汇[27] [28]。对于这一争论，Brown [29]提出了自己的假设，认为完形填空题型所考查的语言能力构念与考生本身的语言能力水平相关。对于语言水平低的考生，完形填空题型可能考查低层次的句内认知加工，因为考生会受到语言水平的限制而无法加工复杂的篇章信息。反之，对于语言水平高的考生，完形填空则是考查较高层次认知加工的有效测试方法。

为了验证 Brown 的假设，McCray & Brunfaut [30]运用眼动技术从考生的认知过程角度研究了集库式完形填空题型考查的构念。通过比较高分考生和低分考生在整体加工(包括文本和词库的加工)、文本加工、任务加工(包括词库加工和与文本加工之间的转换)三个维度上的注视指标。该研究发现，与高分考生相比，低分考生在局部阅读和较低层次的加工上需要更多的认知努力，其注视情况与删除词汇所在的句子语境以及词库中词汇的复杂度有着密切相关。由此，McCray & Brunfaut 得出结论，考生自身的语言水平是影响集库式完形填空题型引发什么样认知过程的重要因素，一定程度上支持了 Brown 的假设。

3.4. 任务模态对考生认知过程的影响

听力理解过程涉及的因素非常复杂，其测试的构念很难定义和描述[31]。大多数研究者认为视觉输入加工是二语听力技能的重要组成部分，因为真实语境中的听力过程是在多模态形式中进行的，听者需同时接收视觉和听觉信息[32]。然而，语言测试学界对于是否将视频材料纳入二语听力测试仍存在争议。有些研究者认为视频材料理应纳入二语听力测试中，这样可以避免构念代表不足，并增加二语听力测试的真实性[33] [34]。然而有些研究者则认为语言测试考查的是考生的语言能力，而非视觉信息理解能力，将视频材料引入二语听力测试中会损害考试的公平性[35]。已有研究通过分析考试分数和考生的有声思维数据探究视觉信息对考生二语听力测试表现的影响，但鲜有研究揭示考生在听力测试过程中观看视频的视觉行为[36] [37]。

为了弥补这一研究空白，Suvorov [37]采用眼动技术记录了 33 名受试在完成基于视频的学术听力测试(The Video-based Academic Listening Test)时的眼动指标，以探究考生在听力测试过程中观看两类视频材料(情境视频和内容视频)时的眼动参与以及考生的眼动数据与考试表现之间的关系。通过配对样本 t 检验，Suvorov 发现被试在观看两类视频材料时的注视比率和总停留时间有显著差异，而停留比率无显著差异；被试的眼动指标和考试分数之间并未呈现显著相关。这意味着被试在考试过程中会与视频材料产生密切互动，且加工不同类型视频的视觉行为有所差异，但视频材料的使用并未对考生的考试表现产生明显影响。同时 Suvorov 强调二语听力测试构念的界定应取决于不同的目标语言使用域[38]，如果视觉信息是目标语言使用领域中不可或缺的一部分，相应的听力测试应引入视频材料。就该研究中的测试而言，视觉信息在所考查的学术听力过程中扮演着重要作用，相应的听力测试应引入视频材料。

3.5. 评分员的认知过程

语言测试认知过程研究的对象除考生外，还有评分员。就写作测试这样的主观性测试而言，评分员信度是衡量测试质量、保障测试结果公平、公正的重要指标。然而，由于个体不同的性格和专业背景，评分员在写作文本评分中较易产生差异。为减少这种差异，提高主观评分的信度，测试开发者尝试过多种方法与措施，其中之一就是评分量表的使用。评分量表在主观评分中被视为引导评分员认知过程的导图[39]。Barkaoui [40]的研究发现评分量表对于评分员决策的影响甚至超过评分员自身的评分经历。在整

体评分中，评分员倾向于仅关注他们自身认为重要的评分标准，而在分项评分中，评分员会关注到评分量表中所有的评分标准[40] [41]。

评分员对保证评分质量和测试公平有着举足轻重的影响。那么，评分员在使用评分量表时经历了怎样的认知过程？评分员在评分过程中会关注评分量表中的哪些标准？Winke & Lim [42]通过眼动技术探究了评分员在使用评分量表对英语作文进行评分决策的认知过程。基于评分员的注视时间和注视次数，研究发现评分员最关注“结构”和“内容”分项标准，最不关注“写作规范”分项标准。Winke & Lim推测，这有可能是因为评分量表的排列顺序而导致的首因效应(the primacy effect)，即最先出现在评分量表中的分项标准最容易被评分员记住，且对评分员的评分影响最大。由此，他们指出写作评分量表的设计对于评分员的评分以及测试构念的阐释都具有重要意义。

4. 不足与反思

眼动技术的应用为语言测试领域认知过程研究提供了一种全新的途径，但相关研究还存在较多局限性。首先，眼动技术有其本身的客观局限性。眼动技术能呈现时间和空间维度上精确的眼球运动数据，帮助研究者有效推测个体的心理认知过程，但不能直接揭示信息加工的生理机制，解释考生的认知过程。因此，相关研究应促进眼动数据与其他来源的数据结合，采用混合式研究方法为揭示语言测试认知过程提供更为充分的证据。此外，在眼动实验中，研究者需要事先对刺激材料进行一定处理，但目前的眼动技术对较大篇幅视觉刺激的追踪敏感度不高，如何真实地呈现语言测试任务也是语言测试眼动研究亟待解决的问题。

其次，现有语言测试眼动研究大多采用非标准的实验设计，并没有准确合理地划分和控制兴趣区，这在一定程度上可能损害眼动数据的准确性，造成研究结果的误差。在眼动数据的选择上，上述研究选用的指标多以注视时间和注视次数为主，比较单一。事实上，可供分析的眼动指标种类非常丰富，研究者可以根据其研究目的综合使用多种眼动指标，验证所推测的认知加工过程和具体的认知负荷，如结合平均注视时间，向前眼跳次数，回视次数，注视位置等整体分析指标从宏观上分析被试阅读的眼动特征，推测被试的认知过程[6]。

最后，眼动技术在语言测试领域的应用研究中，被试的数量有限也是一个显著的局限。现有研究被试的数量都在二三十人左右，或者更少。与语言测试，尤其是大规模、高风险的语言测试考生规模相比，二三十人的样本容量显得十分有限，其研究结果难以概化推广。且现有研究大多止步于对眼动数据的阐释和研究问题的解答，较少深入探究数据结果产生的原因以及对现实语言测试实践的建议。

5. 未来应用展望

眼动技术在语言测试研究中的应用还处于萌芽状态，相关实证研究数量尚不多，但这些创新性尝试为未来的研究提供了宝贵的借鉴。笔者认为在语言测试研究中应用眼动技术，可以从认知过程视角进一步推动构念、效度、考生表现、评分员信度、考试任务设计和开发等研究。

5.1. 构念研究

测试的核心是构念，构念指测量的能力特质。构念既是编写测试任务的基础，也是解释和使用测试分数的依据。明确界定考查的构念是语言测试开发首先需要解决的问题。随着拟测语言能力的变化和考试形式的丰富，语言测试所考查的构念也会随之改变。以往研究对测试构念的定义和验证主要依赖专家判断和考试分数的量化统计分析[18]，很大程度上忽视了考生自身经历的思维过程，并且量化统计数据本身无法形成概念定义[16] [17]。眼动技术从考生的视角出发，记录考生在自然状态下完成测试任务的眼球运动，对于界定语言测试的构念具有重要意义。以听力测试为例，在传统的听力测试中，考生只需要对

听觉信息进行加工处理，听力测试所测量的构念十分明确[18]。然而，视频材料的引入使得听力测试不仅考查考生对听觉信息的加工处理，还涉及考生对非听觉信息的理解能力。因此，为了探究含视频材料的听力测试构念，研究者可以招募与目标考生群体背景相似的被试，采用眼动技术记录被试在传统听力测试和含视频材料的听力测试中的眼动数据，推测考生在完成两种听力测试任务时经历的认知过程，比较分析实际考查的构念。

5.2. 效度验证研究

“除非我们能够证明依照测试结果对考生语言能力所做的推论是有效的，否则无法论证基于考试成绩所做的决策。”[38]。效度一直是语言测试研究的主要议题，认知效度概念在20世纪90年代逐渐得到研究者的认同[43] [44]。该概念强调从心理认知过程的视角补充传统的结果导向型构念效度验证模式。然而，以往的认知过程研究依赖于有声思维和回顾性访谈等质性研究方法，这些研究方法从考生的视角出发，在提供丰富的认知加工过程数据的同时，也因其自身的局限性而受到质疑。有声思维要求考生在答题的同时口头报告思维过程[44] [45]，给考生增加了额外的认知负荷[46]。回顾性访谈这种延迟报告可能掺杂考生对考试过程的失真与过度描述。此外，基于质性方法的过程研究样本容量往往较小，数据分析比较主观，研究结果可推广性不强。

眼动技术能够客观真实地记录语言加工活动中即时的眼球运动，对受试干扰较少，数据客观丰富，为语言测试过程导向型效度验证研究提供了新的可靠数据来源。以阅读测试研究为例，研究者可以通过眼动技术记录考生在考试过程中的眼动数据，通过眼动指标特别是兴趣区内注视时间、回视次数和眼跳距离等指标反映考生的认知加工负荷，利用可视化数据如注视轨迹图和热点图还原考生的作答过程，比较考生在真实环境和考试环境中的认知过程异同，探究构念在实际测试中的实现情况，为测试效度提供强有力的证据。

5.3. 考试表现研究

影响考试表现的因素很多，相关语言测试研究主要围绕三个方面开展：测试程序的特征、考生的答题过程和策略、考生的个体特征[47] [48]。其中，通过探究考生的答题过程和答题策略，研究者可以推测考生测试中具体语言资源的使用、策略使用以及不同层次的认知加工。通过结合即时的眼动数据和考生的作答情况，语言测试学者可以深入分析考生的测试表现，发现考生在考试中遇到的困难和挑战，给语言学习和教学提供建议和反馈。

5.4. 评分员信度研究

对于主观性测试而言，影响测试质量的一大重要因素就是评分员信度。评分量表是测试构念的操作化描述[49]，评分量表的使用是确保写作和翻译等主观性测试评分效度的重要方法之一。通过眼动技术记录评分员在评分过程中阅读评分量表的眼动数据，可以获悉评分员决策时所接收和加工的评分标准，反映评分员对评分量表中各项标准的关注，揭示评分员的评分过程，指导评分员的评分实践。

5.5. 考试任务设计和开发

上世纪90年代，Bachman [47]结合实证研究成果，提出了交际语言能力(Communicative Language Ability)模型。他将交际语言能力定义为结合语言知识和语言使用场景，创造和解释意义的能力。在2001年，《欧洲语言共同参考框架》将交际语言能力作为核心理念进行详细阐述，对欧洲各国甚至世界的语言教学和测评产生了重要影响[50] [51]。对语言测试的启示在于，测试不仅要考查语言形式，还要考查考生在目标情境中语言使用的能力。因此，考试任务特征需要反映真实语言使用任务的特征。然而，单一

模态的测试无法全面考查考生的真实语言交际能力。多媒体和计算机技术的发展使得多模态测试任务以及综合技能测试任务成为了现实，眼动技术能够记录考生加工静态文本、图片、动态网页和视频时的即时眼球运动，揭示考生完成多模态考试任务和综合技能考试任务时的思维过程，为考试任务的设计和开发提供有效反馈，确保测试内容的全面性，增强考试任务的真实性。

6. 结语

虽然眼动技术被广泛应用于认知科学和心理学研究中，并逐渐受到二语研究学者的关注，但在语言测试领域，相关的实证研究数量不多。将眼动技术引入语言测试认知过程研究，可以充分发挥眼动数据客观丰富、真实精确、对受试干扰较小的优势，从全新的视角推动语言测试领域构念、效度验证、考试表现、评分员信度、考试任务设计和开发等重要议题的深入探究。

基金项目

国家社科基金重点项目“基于证据的四六级、雅思、托福考试效度对比研究”(项目编号：14AYY010)。

参考文献

[1] Field, J. (2012) The Cognitive Validity of the Lecture-Based Questions in the IELTS Academic Listening Paper. In: Taylor, L. and Weir, C.J., Eds., *IELTS Collected Papers* 2: *Research in Reading and Listening Assessment*, Cambridge University Press, Cambridge, 391-453.

[2] Purpura, J.E. (2013) Cognition and Language Assessment. In: Kunnan, A.J., Ed., *The Companion to Language Assessment*, Wiley/Blackwell, Hoboken, 1452-1476. https://doi.org/10.1002/9781118411360.wbcla150

[3] Kathy, C. and Pellicer-Sánchez, A. (2016) Using Eye-Tracking in Applied Linguistics and Second Language Research. *Second Language Research*, **32**, 453-467. https://doi.org/10.1177/0267658316637401

[4] Liversedge, S.P., Gilchrist, I.D. and Everling, S. (2011) The Oxford Handbook of Eye-Movement. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780199539789.001.0001

[5] Winke, P.M., Golfroid, A. and Gass, S.M. (2013) Introduction to the Special Issue: Eye-Movement Recordings in Second Language Research. *Studies in Second Language Acquisition*, **35**, 205-212. https://doi.org/10.1017/S027226311200085X

[6] 闫国利, 白学军. 眼动分析技术的基础与应用[M]. 北京: 北京师范大学出版社, 2018.

[7] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and Van de Weijer, J. (2011) Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford.

[8] Rayner, K. Reichle, E.D. and Pollastek, A. (2005) Eye Movement Control in Reading and the E-Z Reader Model. In: Underwood, G., Ed., *Cognitive Processes in Eye Guidance*, Oxford University Press, Oxford, 131-162. https://doi.org/10.1093/acprof:oso/9780198566816.003.0006

[9] Rayner, K. (2009) Eye Movements in Reading: Models and Data. *Journal of Eye Movement Research*, **2**, 1-10.

[10] Van Gog, T. and Jarodoza, H. (2013) Eye-Tracking as a Tool to Study and Enhance Cognitive and Metacognitive Processes in Computer-Based Learning Environments. In: Azevedo, R. and Aleven, V., Eds., *International Handbook of Metacognition and Learning Technologies*, Springer Science + Business Media, New York, 143-156. https://doi.org/10.1007/978-1-4419-5546-3_10

[11] Rayner, K. (1998) Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, **124**, 372-422. https://doi.org/10.1037/0033-2909.124.3.372

[12] Duchowski, A.T. (2002) A Breadth-First Survey of Eye-Tracking Applications. *Behavior Research Methods*, *Instruments*, *and Computers*, **34**, 455-470. https://doi.org/10.3758/BF03195475

[13] 吴迪, 舒华. 眼动技术在阅读研究中的应用[J]. 心理学动态, 2001, 9(4): 319-324.

[14] Just, M.A. and Carpenter, P.A. (1980) A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, **87**, 329-354. https://doi.org/10.1037/0033-295X.87.4.329

[15] 李清华. 语言测试之效度理论发展五十年[J]. 现代外语, 2006, 29(1): 87-95.

[16] Cohen, A.D. (2006) The Coming of Age of Research on Test-Taking Strategies. *Language Assessment Quarterly*, **4**, 307-331. https://doi.org/10.1080/15434300701333129

[17] Weir, C.J. (2005) Language Testing and Validation: An Evidence-Based Approach. Palgrave Macmillan, London.

[18] Bax, S. and Weir, C.J. (2012) Investigating Learners' Cognitive Processes during A Computer-Based CAE Reading Text. In: Banerjee, J., Ed., *Research Notes*: *Issues* 47, Océ Ltd., London, 3-14.

[19] Purpura, J.E. (1997) An Analysis of the Relationships between Test-Takers' Cognitive and Metacognitive Strategy Use and Second Language Test Performance. *Language Learning*, **47**, 289-325. https://doi.org/10.1111/0023-8333.91997009

[20] Purpura, J.E. (1998) Investigating the Effects of Strategy Use and Second Language Test Performance with High- and Low-Ability Test-Takers: A Structural Equation Modelling Approach. *Language Learning*, **15**, 333-379. https://doi.org/10.1177/026553229801500303

[21] Phakiti, A. (2003) A Closer Look at the Relationship of Cognitive and Metacognitive Strategy Use to EFL Reading Achievement Test Performance. *Language Testing*, **20**, 26-56. https://doi.org/10.1191/0265532203lt243oa

[22] Phakiti, A. (2007) Strategic Competence and EFL Reading Test Performance. Peter Lang, New York.

[23] Van Gelderen, A., Schoonen, R., De Glopper, L., Hulstijn, J., Simis, A., Snellings, P. and Stevenson, M. (2004) Linguistic Knowledge, Processing Speed, and Metacognitive Knowledge in First- and Second-Language Reading Comprehension: A Componential Analysis. *Journal of Educational Psychology*, **96**, 19-30. https://doi.org/10.1037/0022-0663.96.1.19

[24] Bax, S. (2013) The Cognitive Processing of Candidates During Reading Tests: Evidence from Eye-Tracking. *Language Testing*, **30**, 441-465. https://doi.org/10.1177/0265532212473244

[25] Bachman, L.F. (1985) Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. *TESOL Quarterly*, **19**, 535-556. https://doi.org/10.2307/3586277

[26] Jonz, J. (1990) Another Turn in the Conversation: What Does Cloze Measure? *TESOL Quarterly*, **24**, 61-83. https://doi.org/10.2307/3586852

[27] Alderson, J.C. (1980) Native and Non-Native Speaker Performance on Cloze Tests. *Language Learning*, **30**, 59-76. https://doi.org/10.1111/j.1467-1770.1980.tb00151.x

[28] Yamashita, J. (2003) Processes of Taking a Gap-Filling Test: Comparison of Skilled and Less Skilled EFL Readers. *Language Testing*, **20**, 267-293. https://doi.org/10.1191/0265532203lt257oa

[29] Brown, J.D. (2013) My Twenty-Five Years of Cloze Testing Research: So What? *International Journal of Language Studies*, **41**, 1-32.

[30] McCray, G. and Brunfaut, T. (2018) Investigating the Construct Measured b Banked Gap-Fill Items: Evidence from Eye-Tracking. *Language Testing*, **35**, 51-73. https://doi.org/10.1177/0265532216677105

[31] Batty, A.O. (2014) A Comparison of Video- and Audio-Mediated Listening Tests with Many-Facet Rasch Modeling and Differential Distractor Functioning. *Language Testing*, **32**, 3-20. https://doi.org/10.1177/0265532214531254

[32] Taylor, L. (2013) Introduction. In: Geranpayeh, A. and Taylor, L., Eds., *Examining Listening*: *Research and Practice in Assessing Second Language Listening*, Cambridge University Press, Cambridge, 1-35.

[33] Ockey, G.J. (2007) Construct Implications of Including Still Image or Video in Computer-Based Listening Tests. *Language Testing*, **24**, 517-537. https://doi.org/10.1177/0265532207080771

[34] Wagner, E. (2010) Test Takers' Interaction with an L2 Video Listening Test. *System*, **38**, 280-291. https://doi.org/10.1016/j.system.2010.01.003

[35] Buck, G. (2001) Assessing Listening. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511732959

[36] Wagner, E. (2007) Are They Watching? Test-Taker Viewing Behavior during L2 Video Listening Test. *Language Learning and Technology*, **11**, 67-86.

[37] Suvorov, R. (2015) The Use of Eye-Tracking in Research on Video-Based Second Language (L2) Listening Assessment: A Comparison of Context Videos and Content Videos. *Language Testing*, **32**, 463-483. https://doi.org/10.1177/0265532214562099

[38] Bachman, L.F. and Palmer, A.S. (1996) Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford University Press, Oxford.

[39] Knoch, U. (2009) Diagnostic Assessment of Writing: A Comparison of Two Rating Scales. *Language Testing*, **26**, 275-304. https://doi.org/10.1177/0265532208101008

[40] Barkaoui, K. (2010) Variability in ESL Essay Rating Processes: The Role of the Rating Scale and Rater Experience. *Language Assessment Quarterly*, **7**, 54-74. https://doi.org/10.1080/15434300903464418

[41] Lumley, T. (2002) Assessment Criteria in A Large-Scale Writing Test: What Do They Really Mean to The Raters? *Language Testing*, **19**, 246-272. https://doi.org/10.1191/0265532202lt230oa

[42] Winke, P. and Lim, H. (2015) ESL Essay Raters' Cognitive Processes in Applying the Jacobs *et al*. Rubric: An Eye-Movement Study. *Assessing Writing*, **25**, 38-54. https://doi.org/10.1016/j.asw.2015.05.002

[43] Baxter, G. and Glaser, R. (1998) Investigating the Cognitive Complexity of Science Assessments. *Educational Measurement*: *Issues and Practices*, **17**, 37-45. https://doi.org/10.1111/j.1745-3992.1998.tb00627.x

[44] Glaser, R. (1991) Expertise and Assessment. In: Wittrock, M.C. and Taylor, L., Eds., *Testing and Cognition*, Prentice Hall, Englewood Cliffs, 17-30.

[45] 郭纯洁. 有声思维在外语教学研究中的应用[M]. 北京: 外语教学与研究出版社, 2015.

[46] Green, A. (1998) Verbal Protocol in Language Testing Research. Cambridge University Press, Cambridge.

[47] Bachman, L.F. (1990) Fundamental Considerations in Language Testing. Oxford University Press, Oxford.

[48] Bachman, L.F. (2000) Modern Language Testing at the Turn of the Century: Assuring That What We Count Counts. *Language Testing*, **17**, 1-42. https://doi.org/10.1177/026553220001700101

[49] 李航. 整体和分项量表的使用对 EFL 作文评分信度的影响[J]. 外语与外语教学, 2013(2): 45-51.

[50] 刘壮, 韩宝成, 阎彤.《欧洲语言共同参考框架》的交际语言能力框架和外语教学理念[J]. 外语教学与研究, 2012, 44(4): 616-623.

[51] 邹申, 张文星, 孔菊芳. 《欧洲语言共同参考框架》在中国: 研究现状和应用展望[J]. 中国外语, 2015(3): 24-31.