第一章
引言

阅读是人类重要的认知活动，是人类获取信息的主要方式。对外语学习者而言，阅读也是提高外语能力的关键途径。阅读能力的提高，能够带动其他能力进步。对阅读能力进行诊断，有助外语学习者认清自身优势与不足，能使教师做出合理决策。当前，高考英语科目一年多考政策已在部分省市实施，若只为考生提供一个“总分”恐难满足高考改革的现实需求，也不符合当代效度验证的理论要求。

本章将从诊断的定义谈起，介绍诊断信息对语言学习的作用，以及运用认知诊断模型从非诊断性测试中提取诊断信息的可行性。分析高考英语改革对诊断信息的需求以及当代效度理论对成绩报告的要求。随后，探讨对阅读能力进行诊断的意义。

1.1 研究背景

1.1.1 诊断信息的作用

“diagnosis”（诊断）一词在《牛津英语词典（第三版）》（*Oxford Dictionary of English 3rd Edition*）中原指医生“通过检查病患症状而对疾病的性质做出的识别和判断”。无论中医还是西医，均需借助一定的手段或工具来诊断病情、对症下药。语言测试借用了这一概念，期望借由测量工具（assessment instrument）“诊断”学习者在语言知识、技能或能力掌握方面的优劣情况，并将这些结果（即诊断信息）反馈师生（Davies et al. 1999：43；刘润清、韩宝成 2000：11；王振亚 2008：56；武尊民等 2011：15）。在语言测试学科，“诊断信息”甚至被视为“学习和测评的接口”（Alderson 2005），其重要性可见一斑。

遗憾的是，虽然诊断信息对教与学有重要意义，专为诊断目的而研发的语言考试却十分罕见。Alderson et al.（2015：2）总结了两方面原因：一、当今的语言测评产业中起“守门”作用的水平测试（proficiency test）处于绝对优势地位，在升学、入职、移民等各领域发挥着主导作用，致使关注学习过程、成本不菲的“诊断测试”少有人问津；二、在诊断测试本应发挥重要作用的校本测评领域，由于语言教师普遍缺少“测评素养”（assessment literacy），对如何编制或选用诊断测试并不熟悉，也缺少利用诊断信息反馈学生由此改进教学的意识或动力。

除上述现实状况外，人们对诊断性测试的定义和编制方法也存在争议。有些专家（如 Alderson 2005：11）认为诊断性测试应该“自成一派”，只适用于低利害、低风险环境，通常只针对“低层次”（low-level）的语言知识。也有专家（如 Bachman 1990：60；Bachman & Palmer 1996：98）认为任何一种语言考试都具备提供诊断信息的“潜力”，对考生的作答情况进行细致入微的分析也能提供有诊断意义的信息。

总体来看，在语言测评领域，对诊断信息的获取大致有两种途径：一种是用专门编制和实施的诊断测试来获取考生在某一具体语言知识、技能或能力方面的作答表现，以此得到诊断信息。例如，DIALANG（Diagnostic Language Test 诊断性语言测试）、DELNA（Diagnostic English Language Needs Assessment 诊断性英语需求测试）、DELTA（Diagnostic English Language Tracking Assessment 诊断性英语追踪测试）。从内容上看，部分测试的题目多少偏重对语法（动词时态、语态）、词汇（词形转换）等“离散的”语言知识进行测量；从设计上看，不少题目在一定程度上受到结构主义语言学和行为主义心理学的影响，如采用选择题形式来测量写作能力，与当代语言测试的主流理念有一定出入。或许由于上述原因，导致这些“诊断测试”未受到语言教师的普遍关注。另一途径是借助认知诊断模型（Cognitive Diagnosis Models，简称 CDMs），来提取非诊断性测试中的“诊断信

息”。认知诊断模型是新一代测量理论的核心，强调对学生的微观心理加工过程的评估，强调关注相同成绩学生的不同认知结构，强调对学生个人的优势与不足的诊断，并为学生提供“颗粒化”（fine-grained）的诊断信息，以促进全面发展（Frederiksen et al. 1993；Leighton & Gierl 2007；Rupp et al. 2010；汪文义、宋丽红 2015；罗照盛 2019；涂冬波等 2019）。

近十几年，国外已有研究者尝试运用认知诊断模型从现有语言水平测试中提取诊断信息、制作成绩报告并反馈师生。例如，美国 SAT 预考利用“规则空间模型”（Rule Space Model）提取考生对阅读能力掌握的情况，并在成绩报告中为考生呈现最需要提高的阅读技能（Buck et al. 1998）。还有研究者（Jang 2005，2009a；Lee & Sawaki 2009a；Kim 2011）运用“融合模型”（Fusion Model）在 TOEFL 考试的阅读、听力和写作测试中提取有关考生语言技能的诊断信息，并制作有诊断意义的成绩报告反馈师生。

1.1.2 高考英语改革的现实需求

高考是高校选拔录取人才的重要手段，是我国现阶段保障社会公平的重要制度。高考不仅对人才选拔的质量有直接影响，对我国基础教育也有强大的导向作用。高考改革一向受到全社会的密切关注。2007 年 5 月在高考恢复 30 周年之际，由教育部考试中心与《中国青年报》联合举行的调查显示：95% 的民众赞同高考制度，但很多人认为高考需要改革（刘海峰 2007）。在谈及高考如何改革时，人们重点关注的是考试时间的选择、考试次数的增加等问题，对考试理论、试题质量、成绩反馈以及考试对教与学的反拨作用等较为“深层”问题却鲜有涉及。

《国家中长期教育改革和发展规划纲要（2010-2020 年）》（以下简称《规划纲要》）明确提出“完善高等教育考试招生制度”“深化考试内容和形式改革，着重考查综合素质和能力”“探索高考部分科目一年多次考试的办法，探索实行社会化考试”等要求。2013 年 11 月《中共中央关于全面深化改革若干重大问题的决定》将《规划纲要》中的

“部分科目”明确为“外语科目”。2014 年 9 月国务院正式出台《关于深化考试招生制度改革的实施意见》，确定高考外语科目一年多考方案，并且提出了“完善成绩报告”的要求。这是新中国成立以来，国家最高层面首次对“成绩报告”提出具体要求。2020 年 6 月，中央全面深化改革委员会第十四次会议审议通过的《深化新时代教育评价改革总体方案》明确提出了“改进结果评价，强化过程评价，探索增值评价，健全综合评价”等要求。高考英语考试不仅要告别“一锤子买卖”的时代，而且成绩报告的形式与内容也将成为改革的重点。

英语是我国高考统考的重要科目，是为高校招生服务的选拔性考试（刘庆思 2008，2017；教育部考试中心 2014：1）。众所周知，作为一项大规模高利害考试，高考英语对高中英语教与学产生了相当程度的负面反拨（亓鲁霞 2004）。考试设计与研究机构虽然无法直接干预教与学，却可以通过改善考试形式与内容、为师生提供有诊断意义的成绩报告等方式对其产生间接影响。伴随考试“形式”的改革（一年多考政策的实施），各考试机构也都在积极探索对考试“内容”进行变革，上至负责全国卷设计与命制工作的教育部考试中心，下至负责自主命题的地方考试院，无不积极探索如何更有效地测量英语运用能力，对高中英语教与学起到正确导向作用。以北京为例，北京教育考试院计划实施一年两考政策外，也专门成立了相关课题组（主要由北京外国语大学专家领衔），负责对高考英语考试的形式与内容进行优化。2016 年 4 月北京教育考试院公布了《北京市关于深化考试招生制度改革的实施方案》，该方案提出听力与笔试分离，一年两考，并对其他考试内容进行了优化。

在考试形式和内容变革的大背景下，若仍只为考生提供一个“总分”，待每次考试后，考生依旧无法了解自己对具体语言知识、技能或能力的掌握情况，教师照旧无法了解自己教授的班级在英语学习的哪些具体方面有所欠缺，无法有的放矢实施“补救性”教学措施。面对分数相同的考生，教师只能“默认”他们具有相同的语言水平，无法真正做到因材施教。果真如此，一年多考政策的实施和考试内容的变

化，或许仍无法起到“以考促学、以考促教”的作用。

考生若能在考后较短的时间内拿到对自身语言能力有诊断意义的成绩报告，报告中不仅呈现考生的作答表现（总分、分项分、排位等），也包含对考生在某一具体能力（如阅读技能）掌握方面的诊断结果（如哪些技能已经掌握、哪些技能尚需提高），有助于认清自身优势与不足，找到正确的学习方向。任课教师若能拿到以任教班级为单位的成绩报告，报告中既呈现班级的作答表现，也包括全班在语言知识、技能或能力上的掌握情况以及分数相同的考生对语言知识或技能运用的不同状态，则更有助于做出正确的教学决策。

近年来，一些参与高考改革的教育测评专家（如雷新勇 2011；葛为民、李金波 2012a，2012b；李金波 2012；刘庆思 2018；李令青等 2019）也曾多次呼吁改进传统的成绩报告方式，提倡为考生个人提供个性化的、对能力掌握有诊断意义的成绩报告；为任课教师提供分数相同考生在能力掌握模式方面的比较信息。遗憾地是，针对高考某一学科（能力）“诊断信息”提取与反馈的研究仍不多见，专门针对高考英语考试进行的诊断信息提取与反馈研究更是罕见。

1.1.3 测试效度验证的理论要求

效度是一个考试中最重要的议题（AREA et al. 2014: 1）。某种意义上，对测试的研究，其实就是对效度的研究（Weir & O’Sullivan 2011）。近 60 年来，人们对效度的认识历经分类观到整体观的演变。语言测试的效度及验证模式虽然有一些“迟滞”，但总体上还是遵从了这一发展脉络（韩宝成、罗凯洲 2013；罗凯洲 2019）。效度整体观特别强调对测试及其测试结果的使用要有正面导向作用。

在 Messick（1988，1989，1996）整体效度概念（Unitary Concept of Validity）中最值得关注的是增添了“社会后果”（social consequence）这一内容，强调测试开发者除了要控制构念代表性不足（construct underrepresentation）以及与构念无关（construct irrelevant）的因素外，还要和测试使用者共同肩负起“举证”责任，了解测试给社会带来的影响。Messick 的效度整体观引起了语言测试学者的关注（Chapelle

& Douglas 1993；Cumming & Berwick 1996），对语言测试效度理论研究有重要的借鉴作用。然而，效度整体观理论太过抽象，一时难以为普通语言测试工作者所接受。Bachman & Palmer（1996）在 Messick（1988，1989）效度整体概念的基础上提出了“测试有用性模式”（model of test usefulness），以此来指导语言测试的开发和效度验证。测试有用性模式由“信度”（reliability）、“构念效度”（construct validity）、“真实性”（authenticity）、“交互性”（interactiveness）、“影响力”（impact）及“可行性”（practicality）等六个质量要素组成（Bachman & Palmer 1996：18）。该模式明确了哪些质量要素需要证据来验证，有一定操作性。但是，质量要素之间的关系并不明确，如证据如何收集、程度如何把握，Bachman 与 Palmer 并未给出令人信服的说明（韩宝成、罗凯洲 2013）。

效度整体观指导下的测试有用性框架也强调了测试的“社会后果”（social consequence）。受其影响，语言测试学科开展了一系列关于测试“反拨作用”（影响力）的研究（如 Alderson & Wall 1993；Shohamy 1993，1996，2001；Cheng 1997，2005；Watanabe 1997；Wall 1999；亓鲁霞 2004；Wall 2005；Hawkey 2006；Wall & Horák 2006；辜向东 2007）。这些研究大都以调查方法为主，回答了测试对教与学的哪些方面产生了影响，并且探讨了负面反拨作用的成因。遗憾的是，上述研究对如何才能产生正面反拨的探讨较少，但研究者普遍相信在保证测试自身质量的前提下，如能为师生提供对语言能力有诊断意义的成绩报告（反馈），引导“用户”合理使用测试、正确看待分数，则有可能放大考试的正面导向作用。

效度整体观虽然使效度理论变得“简约”，突出了构念效度的地位，并且使人们意识到了测试对社会的影响力，但由于缺少可行的验证框架，使效度验证过程仍然像证据罗列。教育与心理测量领域形成一种较为“尴尬”的局面：一方面，在理论界 Messick 的整体观得到普遍认可；另一方面，在效度验证的实践环节，人们仍旧倾向于效度分类观。因为在分类观下，人们至少了解要从哪些方面收集效度证

据。进入上世纪 90 年代后，Kane（1992，2001，2004，2006）提出的基于论证的验证模式（argument-based approach to validation）真正使效度证据的类别得到明确，也间接地“限定”了效度验证的“起点”与“终点”。在效度论证观的大背景下，Bachman（2003）、Bachman & Palmer（2010）提出了“测试使用论证”框架（Assessment Use Argument，简称 AUA），用以指导测试研发与使用。测试使用论证借鉴了 Kane（2002，2006）的论证模式，也包含两个过程，即构建并验证 AUA。

构建 AUA 需要列出关于“后果”（consequences）、“决策”(decisions)、“解释”（interpretations）和“测试成绩”（assessment records）四方面的各种声明（statements），包括“主张”（claims）、“理由”（warrants）和“反驳”（rebuttals），形成一个类似于解释性论证的推理链。上一推理过程的结论（主张），经论证后，就成为一下推理过程的事实（Bachman & Palmer 2010：100）。AUA 的推理链是否成立，需要各类证据支撑。由此，效度验证过程变了一个环环相扣的论证过程。值得注意的是，Bachman 与 Palmer 在阐释 AUA 的主张和理由时，曾多次强调成绩报告在测试产生正面“后果”（beneficial consequence）方面的重要作用；也强调了诊断信息对语言学习者和教师做正确“决策”的重要意义（Bachman & Palmer 2010：196）。这里的诊断信息指通过测试获得的考生在语言能力（某些具体方面）的优势与不足。

在 AUA 的推理链中，最核心的问题莫过于语言能力构念的界定以及测试分数是否能“解释”考生的语言能力（构念）。与上述问题相比，测试是否能提供具有诊断意义的成绩报告指导教与学似乎处于次要地位，但不能否认，它同样是支撑 AUA 推理链不可或缺的证据来源。遗憾的是，不少考试机构在收集效度证据时，更看重统计指标，较少关心测试是否提供有益反馈。

1.1.4 阅读能力的诊断

在现实世界中，尽管科技日新月异，沟通方式愈发多样，但阅读（包括用母语或二语阅读）仍然是人类获取信息的基本手段，也是人类

认识客观世界、培养思维能力、提升道德修养的基本途径。毫无疑问，阅读是人类社会一项十分重要的活动，它使人类进化与发展的宝贵经验得以延续。世界各国都十分重视阅读教学与测评，例如在教育与心理测量领域最为先进的美国，大规模阅读评测活动已开展近半个世纪。创立于20世纪60年代的国家教育进展评估（National Assessment of Educational Progress，简称NAEP）为美国政府合理配置教育资源、制定教育政策提供了可靠依据。NAEP由美国国会授权，在全国范围内重点评测4、8及12年级学生的阅读和数学能力，了解学生的阅读现状及阅读能力变化是NAEP的主要目的。此外，美国“不让一个孩子掉队”教育法案（*No Child Left Behind Act of 2001*）中还明确规定，美国实施的所有中小学测试都应为家长、教师及学生提供诊断信息（成绩报告）。

只考试、不反馈诊断结果在发达国家（尤其是美国）教育界恐怕会被视为是一种不负责任的表现。这里的“诊断”主要指通过考试，对学生某一（学科）知识、技能或能力的掌握情况做出认定、分类、评判后所得到的具体信息，而“反馈”主要指为考生及其他利益攸关者（包括学校、教师、家长等）提供的“成绩报告”。成绩报告中不仅包含了考生的得分，还包含对考生能力的具体分析。在美国，成绩报告的内容往往不是一个简单的原始得分，而是由可比性质的量表分数、分数解读的参照标准、测量误差的说明、改进建议等丰富且实用的信息组成。成绩报告还可依照考生个人、班级、学校、学区的需求而改变具体内容和呈现方式。

在成绩报告中呈现上述信息主要以先进的测量理论为依托。传统的测量理论关注的是测验分数结果或考生能力估计值，关注考生之间水平的相对位置，对测验分数或能力估计值背后所隐藏的认知结构、知识状态无法提供具体信息（罗照盛2012：134）。然而，具有相同分数的考生，可能具有不同的认知结构或知识状态，这些用传统的测量理论模型无法得到解释。不能了解学生的个体差异，即便提供了一定的诊断信息，仍然无法做到有教无类、因材施教。

认知诊断模型（Cognitive Diagnosis Models）能帮助考试研发者

实现对每位考生提供个性化诊断反馈服务的设想，是继经典测量理论、概化理论、项目反应理论之后的新一代测量理论中的核心内容（Mislevy 1993；涂冬波等 2012；辛涛等 2012；汪文义、宋丽红 2015；罗照盛 2019）。近些年来，美国很多考试机构都试图从现有考试中提取认知属性（cognitive attributes），通过认知诊断模型的运用，为考生提供精准的诊断信息。认知诊断模型的运用大致分为三个步骤：Q 矩阵构建、参数估计和模型拟合检验。所谓 Q 矩阵（Q matrix）是由试题和所测构念（测量目标）构成的二维关联矩阵（incidence matrix）。通常用数字 1 表示某小题考到了某个技能（属性）[①]，用 0 表示未考到。对阅读测试而言，Q 矩阵是试题和阅读技能（测量目标）的桥梁。Q 矩阵的质量直接影响参数估计的精准程度，Q 矩阵构建一般以考试大纲中对构念的界定为依据，同时也需要构念维度的证据。

对阅读能力实现认知诊断，就要对阅读的能力维度进行深入分析。大多数关于二语（外语）阅读能力维度的研究都支持“多维假设”，但阅读能力由哪些维度构成，不同的研究环境（研究工具、被试）得到的结论并不一致。

目前，我国各类考试受现实条件所限，通常只为考生提供原始总分或分项成绩，很少为考生、家长及教师提供分数解读的标准，也不会提供个性化的诊断分析。我们的考试理论与技术，乃至教育理念都亟待更新。我国各类考试（如高考、研究生入学考试、公务员考试）都十分重视对阅读能力的考查。以高考为例，无论是全国卷还是自主命题试卷中，阅读试题所占比重最大。例如，2019 年北京高考语文考试中，单对阅读能力考查的试题在总分 150 分占了 90 分；2019 年北京高考英语考试中，与阅读直接或间接有关的试题分值多达 85 分（含语法填空与完形填空）。人们对阅读的重视程度单从高考语文和英语考

① 外语教育领域（包括语言测试学科）和教育与心理测量领域的表述及用词略有不同，本书不刻意区分两大领域的表述习惯。行文强调任务设计（命题）时，主要使用构念（测量目标）等；侧重语言能力时，主要使用技能、能力等；涉及测量模型时，主要使用属性、因子等。

试的分值设置上也可窥探一二。遗憾的是，虽然高考中的语言考试如此重视阅读能力，但受各种条件制约，未能为考生提供详细的（对阅读能力有诊断意义的）成绩报告。伴随高考英语考试形式和内容的改革，师生希望得到的不仅是一个“总分”，还有对具体知识或技能掌握情况有诊断作用的信息。

1.2 研究目的

本研究借助高考英语改革的契机，以英语阅读测试为切入点，在深入探究阅读能力维度的基础上，运用认知诊断融合模型，从阅读试测的原始数据中提取对考生阅读能力有诊断价值的信息（即阅读技能掌握概率），并结合考生作答表现，制作个性化的成绩报告，以期对高中英语教与学产生正面反拨。研究将分为四个阶段逐步实现上述目的（参见图 1.1）：第一阶段为基础研究，重点分析阅读试卷（测量工具）的基本质量指标，也为后续研究提供经典测量理论范式下的数据参照。第二、三阶段为主要研究。其中，第二阶段旨在探究阅读能力的维度，为认知诊断 Q 矩阵的构建提供主要证据；第三阶段旨在运用融合模型提取考生对阅读技能的掌握情况；第四阶段为辅助性研究，旨在运用诊断信息并结合其他内容为师生制作成绩报告。

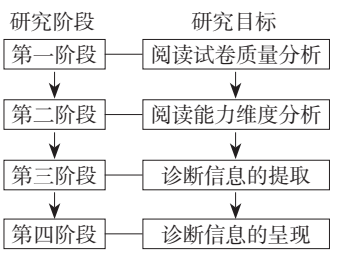

图 1.1 研究阶段与目标

1.3 研究问题

运用认知诊断的融合模型从现有外语阅读水平考试中提取诊断

信息（阅读技能的掌握情况），不啻为一种可靠且高效的手段（Jang 2009a；Sawaki et al. 2009a）。这种手段的使用存在两个预设，一是阅读能力（构念）是可分的、多维的，因为后续的成绩报告中是以阅读能力“子维度”（属性、阅读技能）为纲进行汇报；二是融合模型适合语言测试研究环境、可以用来提取对语言能力有诊断意义的信息。这也分别成为本研究第二、三阶段的重点。此外，由于本研究使用的主要工具为高考英语阅读测试样题（简称“阅读样题”），其基本质量有待进一步验证，故本研究第一阶段把测量工具的质量分析作为研究重点。第一阶段得到的相关指标也能为后续研究（如能力维度分析中的因子负荷、认知诊断和项目反应理论范式下的试题难度等）提供参照。提取诊断信息的最终目的是为了“反馈”，反馈的主要形式是成绩报告，由于成绩报告不能只呈现“诊断信息”，还得包含“考生分数、排位、对错情况、试题难度”等信息，这些信息与“诊断信息”一道，在辅助考生认清自身优势与不足方面有协同作用，所以本研究也会关注如何呈现上述信息，这成为第四阶段的研究重点。

本研究各阶段待回答的问题如下：

第一阶段：阅读测试样题（测量工具）的质量是否达标？

该问题将被分解为若干操作性问题：整份阅读样题的质量参数（如信度系数、标准误）是否符合经典测量理论的标准？小题的质量参数（如难度、区分度及标准差）是否符合经典测量理论的标准？阅读样题的文本材料有何特征？阅读样题的试题类型有何特征？考生对阅读样题的总体设计持何态度？英语教师对阅读样题的总体设计持何态度？

第二阶段：（本研究环境下的）英语阅读能力由哪些维度构成？

该问题将被分解为若干操作性问题：《普通高等学校招生全国统

一考试北京卷英语考试说明（讨论稿）》[①]（以下简称《考试说明（讨论稿）》）对阅读能力的规定是什么，命题人如何界定阅读试题与测量目标的关系？试卷是否拒绝单维性假设？如果拒绝，起主导作用的维度会不会是题组维度（效应）？如果不是，试卷可能包含哪几个维度？此外，考生作答试题所用的阅读技能（策略）是否与命题人（测量目标，即阅读技能）的预期相符？专家对阅读样题的结构（阅读能力维度）的判断是否与命题人一致？如果不一致，专家对阅读能力维度有何假设？基于结构方程模型的验证性因子分析对哪种假设模型更为支持？

第三阶段：如何运用融合模型提取对考生阅读能力有诊断意义的信息？

该问题将被分解为若干操作性问题：实现融合模型操作的 Q 矩阵如何构建？各参数的马尔可夫链是否收敛？参数估计结果（如 π_i^*、r_{ia}^*、c_i 值）是否达标？融合模型与数据是否拟合？诊断结果出现“异常”的个案该如何解释？

第四阶段：如何为师生制作包含诊断信息的成绩报告？

该问题将被分解为若干操作性问题：哪些信息将在成绩报告中呈现，其参照依据是什么？如何为师生在成绩报告中呈现上述信息？师生如何评价成绩报告的内容与形式？

1.4 本书结构

本书共分八章。

第一章为“引言”，重点介绍研究背景、研究目的及研究问题。

第二章为“文献综述”，将回顾阅读能力维度的相关研究，以及心理测量理论在语言测试中的应用，重点评述认知诊断融合模型在提取（语言能力）诊断信息方面的研究。

第三章为“研究方法”，首先对本研究四个阶段将采用的量化和

① 《普通高等学校招生全国统一考试北京卷英语考试说明（讨论稿）》由高考英语考试改革研究课题组提供。

质性研究方法进行介绍；然后，对研究参与者、研究工具等进行了较为详细的描述；最后，对数据收集与分析步骤进行梳理和解释。

第四章至第七章为各研究阶段的结果分析与讨论。

第四章为“阅读测试质量分析”，将汇报使用经典测量理论指标对试卷质量分析的结果，分析试卷（文本与题型）的特征，并呈现师生对于试题质量的看法。

第五章为“阅读能力维度分析”，将从多个角度（命题人、专家等）形成关于本研究环境下对英语阅读能力维度的假设，并使用基于结构方程模型的验证性因子分析方法对各假设进行验证，最后得出拟合程度最好的假设。

第六章为“诊断信息的提取”，将依照上阶段关于阅读能力维度的分析结果以及阅读测试答题过程有声思维的分析结果构建 Q 矩阵；然后进行多轮参数估计分析，最后，呈现模型拟合结果，确定提取到的诊断信息的可靠性。

第七章为“诊断信息的呈现”，将对成绩反馈信息的主要内容及依据、呈现形式进行汇报，并分析了师生对诊断信息（成绩报告）的看法。

第八章为“结论”，将回顾各阶段的研究发现，并对本研究的理论和实践价值加以总结；最后，分析研究不足，并对后续研究进行展望。

第二章
文献综述

本研究采用认知诊断模型从“非诊断性测试”（高考英语阅读测试）中提取了对考生英语阅读能力有诊断意义的信息。这里有两大问题有待探讨：一、外语（二语）阅读能力的维度（即阅读能力由哪些方面构成），这直接关系到Q矩阵的构建，是实现认知诊断的基础；二、选用合适的认知诊断模型进行参数估计并进行拟合检验，是提取诊断信息的关键。对阅读测试而言，阅读能力维度也是界定阅读测试构念的重要依据，对阅读能力维度的验证，从某种程度上来说，其实也是在对阅读测试进行“构念效度验证”，是一个问题的两个侧面。

本章将从构念的界定谈起，回顾以往研究中对母语和外语（二语）阅读能力维度（构成）的研究，总结二语阅读构念界定的主要思路。然后，介绍认知诊断模型诞生的背景，分析如何对模型进行选择。最后，解释融合模型的数学表达并回顾融合模型在语言测试中的应用。

2.1 阅读能力研究

2.1.1 构念与构念界定

构念（construct）指“一项测试要测量的概念或特征”（AERA et al. 2014：173）。这是2014年版《教育与心理测验标准》（*Standards for Educational and Psychological Testing*）（以下简称《标准》）为“构念”下的定义。与老版本《标准》相比，新版《标准》已经赋予构念

核心地位，其中最直接的体现就是把传统的“构念效度”直接定义为效度，而其他“种类的效度”（如内容效度等）都作为效度（即构念效度）的证据来源。

在教育与心理测量领域，被测量的对象可能是智商（intelligence）、学能（aptitude）、动机（motivation）、问题解决技能（problem-solving skills）、思辨能力（critical thinking ability）、阅读能力（reading ability）等任何抽象的理论概念（或假设）。这些概念各自有种类繁多理论模型，对于一次测试而言，不可能对理论模型中的所有要素都进行测量，测试目的不同，测试内容也会相应变化。确定一次测试要测量的内容就是在进行所谓的构念界定（construct definition）。从这个角度来看，对构念进行界定好似对理论模型的要素进行“抽样”。构念界定除依照理论外，有时也参照实际需要。例如，对飞行员的执业资格的考查，不仅要涉及飞行理论常识，还得涉及实际飞行能力。所以，构念界定也不一定单纯依靠具体理论模型。无论对考试内容的甄选，还是对实际（作答）表现的考查，一般都离不开相关学科基础理论的支撑。

构念在不同的学科或语境中，时而被称为特质（trait）、属性（attribute）、测量目标（assessment objectives），时而指代知识（knowledge）、技能（skill）或能力（ability）。叫法虽不同，差异并不大，都是不可直接观测（unobservable）的抽象概念或特征。对绝大多数测试而言，构念的重要性毋庸质疑，既是测试任务编制的基础，又是成绩解释与使用的依据。人们无法给构念直接赋分，必须透过考生完成某项任务的外显行为表现来打分（Raykov & Marcoulides 2011：2），依照所得分数对考生掌握构念的程度（乃至构念所代表的某种“特质”“技能”或“能力”）进行推测。

可想而知，如果构念在被界定的那一刻就出现问题，不仅任务编制会受到影响，成绩解读的质量也会出现偏差。一般来讲，心理学中的某个理论构念通常都是可分割的概念，但是，人们能够测量的仅仅是理论构念的某些“侧面”而已，是构念的具体表征。在语言测试中，除了从某种有关语言能力的理论出发界定的构念外，有时也从“实际”

出发，如直接把完成目标语言使用域“任务”的表现（performance）作为“构念”。综上，对阅读能力进行测量，需要关注以往研究对阅读本质的看法、对阅读模式及能力的探究，以及对阅读构念的界定方式。

2.1.2 母语阅读模式与能力

阅读是人类最常见的认知活动。阅读的基本目的是在获取视觉文字信息的基础上构建文本意义（Rayner et al. 2012：1）。以往对阅读（理解）能力的研究（Urquhart & Weir 1998；Alderson 2000；Koda 2005；Khalifa & Weir 2009；Weir 2013），主要形成两种观点：一种把阅读（理解）能力视为读者与文本互动的结果，另一种则视为互动的过程。结果观认为只有当文本内容被记忆，理解才会发生。所以，对阅读能力的测量主要通过衡量受试阅读文本后能记住多少内容。由于无法严格区分理解与记忆，导致测试结果难以解读，也就无法为教学提供有价值的反馈信息。过程观则认为工作记忆只为文本信息构建提供空间，而文本信息如何构建、存储以及如何与背景知识互动后形成意义，才是真正的阅读理解过程。过程观对阅读能力的测量通常要求受试在限定的时间内完成涵盖各层次认知加工要求的阅读任务，以此捕捉实时加工过程，但这种方法可能对受试造成某种程度的干扰。

人们对阅读过程的研究主要形成了三类模式（Alderson 2000：16；Grabe & Stoller 2002：31，2011：25；程立 2010），即自下而上模式（bottom-up models）（Gough 1972）、自上而下模式（top-down models）（Goodman 1968，1976）和互动模式（interactive models）（Rumelhart 1977；Stanovich 1980）。自下而上模式好似一个线性模型，读者首先通过图像刺激来识别字母；其次，把字母转化成声音，由此实现从字素到音素的表征转换；然后，把音素转换成为单词；再由单词扩展成句子；句子组成段落；段落构成篇章。总之，在阅读的过程中，后者以前者为基础，逐级识别，最后构建语篇意义。Alderson（2000：17）认为此模式使人不由得与当时盛行的行为主义（心理学）联系在一起。行为主义往往忽视“人”（读者）在阅读过程中的能动作用。

自上而下模式则强调读者已有知识（图式）在文本理解过程中的作用。自上而下模式认为理解不一定是在对每个单词解码后才会发生。在阅读之前，读者就已经有了一定的图式（schema），如背景及语言知识，阅读过程就是激活已有图式、构建图式与所读文本间关联的过程，也是一种“心理语言的猜测游戏”（Goodman 1967）。这种模式虽然强调了读者的能动作用，但仍饱受争议。读者在阅读文本时并非总在猜测或验证，只要是阅读，多少都会使用些背景知识，但高水平的读者反而较少依赖背景知识（Urquhart & Weir 1998：44）。自下而上和自上而下模式都将阅读视作单方向的线性理解过程，因此，它们对阅读本质的解释都存在局限性（程立 2010）。互动模式则认为阅读过程在本质上双向的，有经验的读者能根据需要交替采用两种处理模式。Grabe（1991）认为所谓互动模式，应包含两种互动关系，一是读者与文本间的互动，二是各种阅读技能间的互动。阅读技能之间不是线性关系，而是平行关系。

如何界定母语阅读能力，人们主要形成了两类观点。一种观点认为阅读（理解）能力是个整体，不可分。持此观点的研究者（Goodman 1969；Thorndike 1973）认为阅读能力是整体获得的，对阅读能力进行划分并依此进行教学与测试都违背了这一自然规律。一些研究（Zwick 1987；Rost 1993）采用探索性因子分析方法来证明阅读能力是单维的，但这些研究的对象全都是母语者，而且只采用了探索性因子分析的方法，未能提供多角度证据。另一种观点认为阅读能力可分，例如，Berkoff（1979）和 Carver（1992）则持二分观点，认为阅读能力可分为一般阅读技能和词汇能力。Lennon（1962）与 Carroll（1993）持多分观点。Lennon（1962）认为阅读能力包括四个维度可以测量：词汇知识、理解明示信息、理解隐含意义和赏析（appreciation）。Carroll（1993）则在分析 30 余项研究的基础上，提炼了四个可测量的维度：一般阅读能力、具体阅读能力、解码能力（decoding）和速度（speed）。在母语阅读能力维度的研究中，当进行理论探讨时，研究者往往能列出少则两三个，多至几十种的阅读（技

能）维度，但实证研究结果往往只能支持少数几个能力维度。一些测评专家（如 Henning 1992；McNamara 1996）曾指出心理学（理论上）能够区分的维度在心理测量学上并不一定是多维的。究其原因，可能是统计方法的局限。以 Berkoff（1979）和 Carver（1992）的实证研究为例，采用的都是探索性因子分析的方式，这种方式便于操作，研究者假定每个观测变量（如试题得分）都与某个因子（如阅读能力的某一维度）匹配，只通过因子负荷推断实测数据的因子结构。现在，人们更倾向于验证性因子分析，即在“先验信息”（理论假设）的前提下，探究因子结构是否符合理论假设。实际操作时，探索性和验证性因子分析往往先后使用，也就是说，若没有恰当的“先验信息”帮助研究者形成假设，那就先做探索性分析，然后依照其结果再做验证性分析。

2.1.3 二语阅读模式与能力

二语（外语）阅读与母语阅读存在诸多不同，Grabe & Stoller（2011：34）将其总结为三大方面：一是语言和加工过程的差异，如词汇、语法、语篇知识不同；二是个体和经历的差异，如母语读写能力、阅读动机不同；三是社会文化和教育机构的差异，如成长环境不同。

对二语阅读模式的探究，人们大致形成两种观点。一种基本遵从（母语）阅读“自下而上”模式，不同之处在于强调“读者”在阅读过程中的能动作用（如策略使用）。Khalifa & Weir（2009）提出的“阅读（认知加工过程）模式”是对第一种模式较好的诠释。该模式虽然是在讨论二语测试的阅读构念时提出的，但并未刻意强调母语阅读与二语阅读的区别。该模式汲取了认知心理学对阅读研究的精华，意在指导阅读测试如何对阅读心理加工过程进行测量。Khalifa & Weir（2009）和 Weir（2013）认为阅读加工存在不同的层次（图 2.1 中间自下而上的层次），理解的效果受（图中左列）元认知策略和（右列）各种知识掌握的影响。其中，语言知识（词汇、句法等）发生在较低层。这一模式重点突出了阅读文本时的认知加工过程，如解码印刷符号（decoding）、构建文本信息（text-information building）、建构读者

模式（reader-model construction）等（Koda 2013），但阅读理解到底是不是一个自下而上的加工过程一直存有争议，该模式的科学性有待进一步验证。

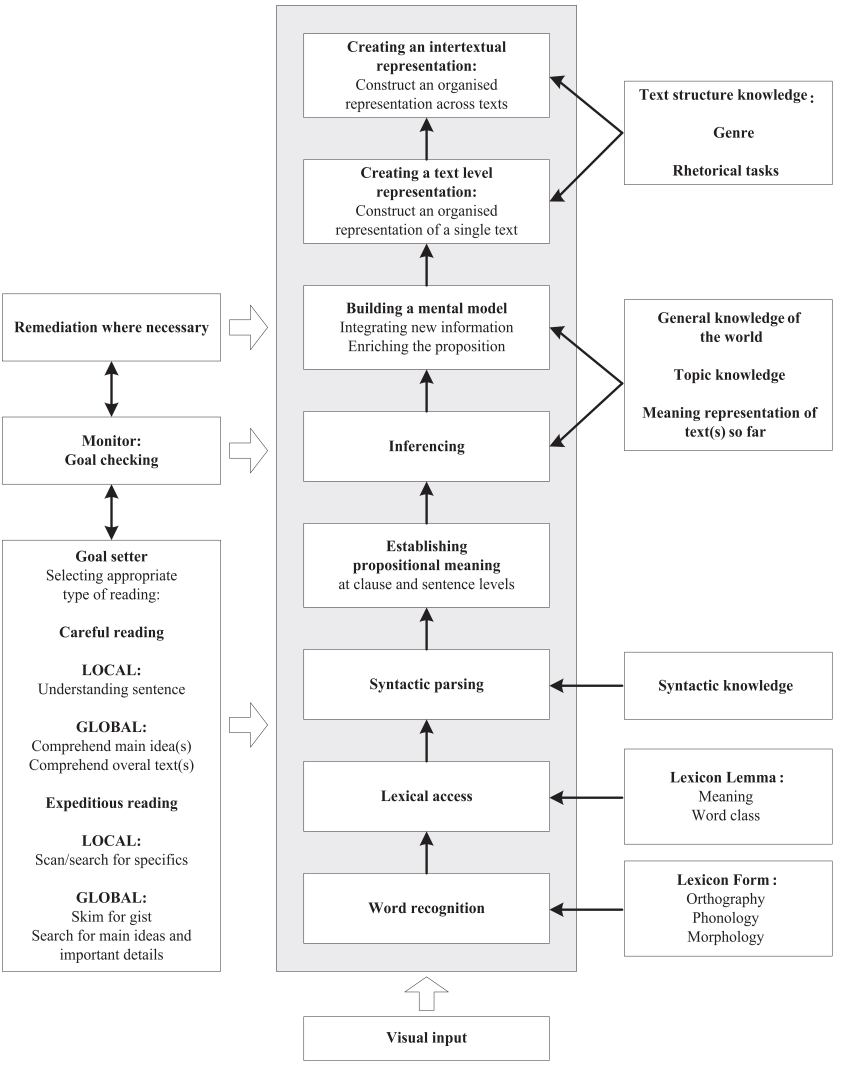

图 2.1　阅读（认知加工）模式（基于 Khalifa & Weir 2009：43）

另外一种观点以 Bernhardt（2005）提出的“二语阅读补偿模式”

（Compensatory Model）为代表（图 2.2）。“二语阅读补偿模式”从心理语言学视角阐释了阅读（理解）能力的发展过程。补偿模式由三部分组成：母语阅读能力、二语知识，以及无法解释的变量“因素”（variance）。Bernhardt（2005，2011）根据以往的实证研究结果，认为母语阅读能力可以解释二语阅读能力的 20%；而二语知识的掌握情况可以解释 30%；剩下的 50% 是不容易确定的影响因素，如策略、背景知识、动机等。所谓“补偿”是指掌握情况较好的知识或能力可弥补较薄弱的部分。“补偿”可以发生在每个部分内部，也可以发生在部分之间。如背景知识可能对二语知识有补偿作用；较高的母语阅读能力可以一定程度上“辅助”尚不完善的二语阅读能力；对二语词汇的掌握可以使读者少犯句法错误等（Bernhardt 2005）。补偿模型中的三部分，并不是简单的累加关系，随着时间的推移，语言水平或其他因素的变化，各组成部分所占比例也会相应改变。可惜这一模式对二语阅读测试的研发缺少指导意义，仍然无法帮助测试研发人员明确二语阅读应该“考什么”“怎么考”。Grabe（2009：104）曾批评这一“描述性质”的模型过于笼统，没有说清楚阅读能力的组成要素，对阅读能力的发展缺少解释力。

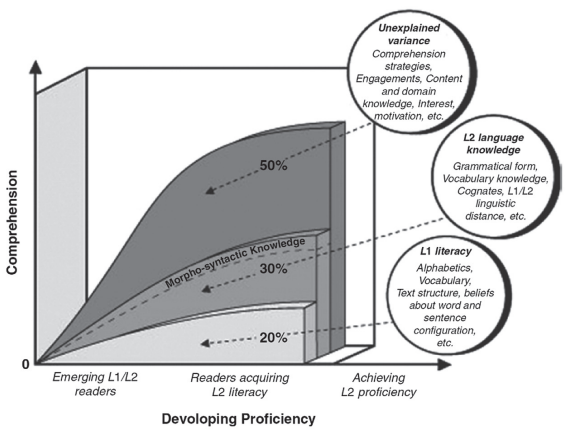

图 2.2　二语阅读补偿模式（Bernhardt 2011：38）

在二语阅读研究中，人们也经常关注二语知识与二语阅读能力的

关系（Alderson & Urquhart 1984）。对二语知识的作用大致形成两种主要观点。第一种以 Alderson（2000：38）为代表，认为二语读者需具备一定的二语知识后，母语的阅读技能才能向二语“迁移”，也就是说在跨越“语言阀限”后，母语阅读能力与二语阅读能力趋同。这也是语言知识阈限假设（linguistic threshold hypothesis）的基本要义。一些研究（Bossers 1991；Bernhardt & Kamil 1995；Pichette et al. 2003）表明，相比基本阅读技能（可由母语“迁移”到二语），二语知识对二语阅读能力也有较大贡献。然而，语言知识阀限也不是“绝对的”，可能因阅读（测试）任务不同（文本话题、文本语言、背景知识、任务类型）而发生变化（Alderson 2000：39）。Khalifa & Weir（2009）的“阅读（认知加工过程）模型”对语言知识的作用持相似看法。另外一种观点在“二语阅读补偿模式”（Bernhardt 2005）得到充分体现，认为二语知识与母语阅读能力以及其他因素（背景知识、动机等）共同构成二语阅读能力。

对二语阅读能力的测评，单靠解读与阐释阅读过程、模式，仍然无法有效确定测量目标。诚然，二语知识对二语阅读理解起到了至关重要的作用，但是大部分二语阅读测试（如 TOEFL iBT、IELTS、PTE、MELAB）通常要测量的是阅读能力中与阅读技能直接相关的构念，而不是语言知识。如前所述，不具备一定的二语知识，考生根本无法完成阅读任务。其实，测量阅读技能的同时，语言知识也间接地得到测量。与母语能力维度（成分）的研究不同，大部分关于二语阅读能力（成分）的研究（Carr & Levy 1990；Anderson et al. 1991；Grabe 1991，2000，2009；Lumley 1993；Bachman et al. 1996；Weir et al. 2000；Hudson 2007；Koda 2007；Song 2008）都支持能力多分的观点。上述研究的结果对二语阅读能力由哪些维度（成分）构成并未达成一致，如何界定阅读测试构念一直存在争议（参见 2.1.5 小节有关 CFA 与二语阅读能力维度的阐述）。

2.1.4 二语阅读测试构念界定

依照不同的考试目的、用途，以及对阅读能力本质的看法，二

语阅读测试的构念界定大致形成三个视角：阅读任务视角、阅读目的视角和阅读技能视角（Enright et al. 2000；Khalifa & Weir 2009：34；Grabe & Jiang 2014）。

一、阅读任务视角（reading task perspective）。阅读任务视角选取现实世界中真实、典型的阅读任务来界定构念，任务被看作是评价学习者语言运用能力的有效工具。某些专门用途语言测试在设计之初就认定了目标语言使用域（target language use domain）中的阅读任务类型，如阅读病例、给药说明、药物标签、转诊报告等，并以此作为试题命制的依据。根据考生完成任务的表现来推断其语言运用能力。Bachman & Palmer（2010：219）似乎不太赞同以这种视角来界定语言测试构念，认为任务型语言测试的主要目的是“预测”未来表现，不能对考生的语言能力进行有效推断。

二、阅读目的视角（reading purpose perspective）。阅读目的视角是用读者阅读文本的目的来界定构念的方式。Grabe & Jiang（2014）认为用该视角来界定构念是最直截了当的方式。TOEFL iBT 阅读测试就使用这种方式来界定构念，包括寻找信息（reading to find information）、基本理解（reading for basic comprehension）、领会语篇结构（reading to learn）（Enright et al. 2000）。

三、阅读技能视角（reading subskills perspective）。此视角可以根据阅读理论、教学大纲或需求分析结果来界定阅读测试构念。例如，识别事实性细节、理解主旨大意等。我国的外语阅读考试（如 CET 考试及高考英语考试）大都采用此视角。以高考英语北京卷阅读测试构念为例，包括“考查考生阅读理解不同文体语篇的能力。考生应能：1）理解归纳语篇的主旨要义；2）获取事实性信息；3）利用上下文推测词义；4）推断隐含意义；5）理解作者意图、观点和态度；6）理解句子、段落之间的逻辑关系；7）理解语篇的基本结构”（北京教育考试院 2014：91）。美国的 MELAB 考试、PTE 考试等也采用此类视角。此外，虽然剑桥系列英语考试（如 KET、PET、FCE、CAE、CPE）及 IELTS 考试都宣称使用 Khalifa & Weir（2009）的阅读认知加工过

程思路来界定构念和任务编制，但从上述考试官方手册（Cambridge ESOL 2007；Cullen et al. 2014）以及效度验证文献汇编（Taylor & Weir 2012）中都可以发现“技能视角”的影子。其实，TOEFL iBT 阅读测试宣称采用阅读目的视角指导构念界定，但在某些研究中仍然会把托福阅读的九类设问方式与具体的阅读技能相对应（Jang 2005，2009a）。

Bachman & Palmer（2010：220-221）认为以语言技能（如听、说、读、写）作为“构念”可能会给评分和分数解读造成一定麻烦，主张用其提出的交际语言能力模型中的语言能力成分（语言知识和策略能力）来界定构念。其实，考生是否具备语言知识和能否运用语言知识是两回事，至少在 Bachman & Palmer 的著作中，没有明确二者的区别。在主流语言测试中，只用语言知识加策略来界定构建的方式仍很少见，其可行性有待进一步研究。

当前，国际上大型英语考试（如 IELTS、TOEFL iBT、PTE）仍以“语言技能”（听、说、读、写）方式对语言能力进行考查，并分项汇报语言技能的成绩。需要指出的是，虽然这些测试冠以“听、说、读、写”之名，但“听”中有“写”、“说”中有“读”、“写”中有“读”甚至还有“听”。这些考试的构念界定绝非简单依照“技能 + 成分”（Lado 1961；Carroll 1961，1968；Davies 1977），更多的是以交际语言能力（Canale & Swain 1980；Palmer et al. 1981；Canale 1983；Bachman 1990；Bachman & Palmer 1996；Bachman & Palmer 2010）为内核。此时的“听、说、读、写”或任何语言技能的组合（如读写结合等）都被看成实现语言能力构念的“任务”或“活动”而已。可以预见在很长时间内，以“技能”为纲进行语言能力构念界定还将占主导地位，但“技能”（如阅读技能）[①] 的内涵，将以认知心理学和语言习得等学科的最新研究成果为依据。

① 本书遵循语言测试学科的表述习惯，姑且将技能视为能力的组成部分。行文涉及认知诊断融合模型（尤其是 Q 矩阵构建）时，将其称为（技能）属性或属性。

2.1.5 二语阅读能力维度（及构念效度）的验证方法

阅读能力维度、阅读测试构念效度（阅读测试试卷维度）其实是同一事物的不同侧面。“维度”是无法直接测量的“潜变量”（即因子），要通过“观测变量”（如试题、试卷）来探究其结构。对于“维度”的研究（如 2.1.2~2.1.3 小节所述）大都采用因子分析方式。

因子分析

因子分析的主要原理是通过对试题的相关系数矩阵的内部结构进行分析，把高度相关性的试题按某种规则分成亚组，每一亚组都有一个公共因子（common factor），公共因子便代表了参与计算的整份试卷（测量工具）的基本结构。因子分析分为探索性因子分析（Exploratory Factor Analysis，EFA）和验证性因子分析（Confirmatory Factor Analysis，CFA）。探索性因子分析的主要目的是确定试题的潜在因子（构念），而验证性因子分析的主要目的在于验证理论假说（因子及因子关系）是否能得到实测数据的支持。从语言测试的角度看，验证性因子分析一方面可以验证构念（因子）之间的关系，另一方面就是验证试题是否考了既定的构念。

探索性因子分析的缺憾在于没有理论假设前提，仅从数据出发（data-driven），是一种典型的后验分析（posterior analysis），所得结果完全受到数据的制约，加之过于严苛的条件限制，越发不能满足研究需要。探索性因子分析只关心采用不同方法（试题）对一种能力（构念）进行测量后，所得结果是否具有较高相关性。换言之，探索性因子分析通常只能证明某些试题（或一项考试）考了一种或多种特定的能力（构念），却无法证明考试结果未受其他能力（构念）或测试方法（methods）的影响，更不可能明确这些能力（构念）到底是什么。为克服上述不足，人们尝试采用“多质多法”（Multi-Traits Multi-Methods，MTMM）方式做验证性因子分析。所谓多质多法是指采用两种（或以上）的测试方法测量同一构念（特质），并采用同一方法测量不同构念的研究设计。研究者通过多质多法得到一个相关矩阵。理论上，矩阵中同一方法对同一构念测量后的信度系数和采用不同方法对

同一构念测量得到的相关系数都应越高越好；而用同一方法对不同构念测量后的相关程度应该低于前两类系数。

多质多法较探索性因子分析已有较大进步。首先，研究者进行多质多法设计之初就对测试要考查的构念形成了较为明确的判断；其次，多质多法同时考虑了测量结果的“收敛”（高相关）及“区分”（低相关），提供了更有说服力的效度证据。然而，语言测试中的多质多法研究重点要看“区分”的程度，但“区分”多大才算具有显著意义，人们并没有达成共识（Bachman 2004：283）。

如今，虽然越来越多的语言测试研究者更倾向于使用基于结构方程模型的验证性因子分析方式，探究能力维度、验证构念效度。大多数研究者不再采取对立的观点看待探索性及验证性因子分析，更多希望综合二者的长处。这样的观念和实践操作，得益于结构方程模型（Structural Equation Modeling，SEM）及其配套软件的使用。结构方程模型是集因子分析、路径分析（Path Analysis）和线性回归（Linear Regression）为一体的数学模型，探索性因子分析、验证性因子分析（包括多质多法）通常被视为结构方程模型中的特例（许宏晨 2019）。这些分析方法有助于确定试题的同质性，以及构念的辨识度（Bachman 2004：283；Ockey 2014：1224）。当关注潜变量的关系时，实际上就在探究“因子”间的结构（如阅读能力维度的构成），这种关系在 SEM 中被称为“结构模型”（structure model）（韩宝成 2006：78；吴明隆 2013：14）。当关注潜变量和观测变量（阅读试题）的关系时，实际上就是在探究构念效度（如阅读试卷自身的维度）。当然，基于 SEM 的 CFA 分析结果只能算作构念效度证据之一。潜变量和观测变量这种关系的探讨，在 SEM 中被称为“测量模型”（measurement model）（韩宝成 2006：78；吴明隆 2013：14）。

验证性因子分析在语言测试中的应用

CFA 常被视为 SEM 的亚模型。CFA 的优势之一就是直接检验考试设计者所界定的构念是否符合预期。在语言测试学科，人们通常根据语言学、应用语言学、二语习得、心理学等学科或教学大纲、实际

需求等确定要考的内容（即构念）。然后，依照构念命题。因此，对语言测试工作者而言，CFA 比传统的 EFA 或许更有应用价值。

CFA 方法在语言测试中的应用主要分为两大方面：一是探究能力结构（或构念效度）（Schoonen et al. 2003；Lee 2005；Sawaki et al. 2008；In'nami & Koizumi 2012）；二是验证策略使用或测试方法（methods）对成绩的影响（Kunnan 1995；Purpura 1997，1998；Llosa 2007；Sawaki 2007）。

在我国，吴红云和刘润清（2004）运用它研究写作元认知能力与写作成绩之间的关系。韩宝成（2006）向国内语言测试同行介绍了 SEM 在语言测试中的应用。邹申和彭康洲（2009）运用专业八级考试实测数据对人文知识试题的构念（社会文化知识、文学知识、语言学知识）做了验证性因子分析，结果表明二阶单因子模型能够揭示人文知识构念与试题的关系。彭康洲和邹申（2012）再次使用验证性因子分析方法分析了英语专业四级考试语法词汇试题对构念（语言、词汇）贡献和影响，结果显示，考试数据（绝大部分试题）能够与模型拟合，语法以及词汇试题分别具有较好的单维性特征，语言与词汇属于两个高度相关但又相互独立的构念。

验证性因子分析与二语阅读能力维度

近年来，使用 CFA 方法探究二语阅读能力维度的研究大都支持多维假设。Sawaki et al.（2008，2009）对托福网考（TOEFL iBT）试卷进行了 CFA 分析，结果表明 TOEFL iBT 考试可以分成“听、说、读、写”四个组成部分，为成绩分项汇报提供了证据。这项研究对于阅读理解部分单独做验证性因子分析时，结果支持阅读能力单维性假设（unidimensional hypothesis），不支持考试设计者认定的三“因子”假设。考试规范中的三个因子（阅读理解构念）包括基本理解（basic comprehension）、推断（inference）和领会语篇结构（reading to learn），其中领会语篇结构是新加入的考试构念，旨在测量更高层次的阅读理解能力，通常需要考生理解全文的基础上比较主次观点或勾勒全篇概要（Pearlman 2008：244）。

Song（2008）使用CFA方法为一项针对留学生英语理解能力的（包括阅读和听力）测试进行分析。这项考试由加州大学洛杉矶分校应用语言学系开发，阅读部分重点测量“理解主旨大意”、“推断隐含意义”、“理解主次细节”三种技能，题型包括简短回答及文章概要填充等。结果表明在该研究环境下二阶三因子模型（阅读理解能力由上述三种技能组成）与样本数据非常吻合，由此为阅读能力可分性假设提供了证据。

孔文（2009）对英语专业四级考试阅读理解试题的因子结构进行了探究。首先，依照考纲并结合专家判断认定出四个因子，分别是“字面理解”“信息重组或释义”“推理”以及“评判”；然后，构建了三个不同的结构模型，分别是单因子模型（TEM4阅读能力）、双因子模型（由基本字面理解能力和复杂阅读理解能力组成），以及四因子模型（由上述从考纲和专家判断推导的四个因子组成）。结果表明四因子模型被拒绝，双因子模型较单因子模型有更好的拟合表现。值得注意的是，双因子模型中阅读能力的二分成分相关度高达0.83，虽然两个成分在模型中可以相互独立，但相关性甚高。其中，基本理解能力能解释的最小方差比例只有7%，最大方差比例仅为21%，复杂阅读能力能解释的最小方差比例只有10%，最大方差比例仅为25%，这或许可以说明，还有其他潜在因素影响二语阅读能力可能。

构念在心理学上（理论上）可以区分，但却不一定能够得到心理测量学的支持。究其原因，可能是统计方法和样本选用的问题。有些研究者使用探索性因子分析（如Swinton & Powers 1980；Manning 1987），有的则用验证性因子分析（如Sticker et al. 2005；Sawaki et al. 2009b；In’nami & Koizumi 2012），有的使用题目打包（item parcel）成绩作为分析数据（如In’nami & Koizumi 2012），有的则使用小题（individual item）成绩作为分析数据（Sawaki et al. 2009b）。分析方法不同、数据差异以及考生群体变化都有可能影响研究结果（Sasaki 1996）。需要注意的是成绩打包处理与小题单独计分存在较大差异，小题单独计分时，依照理论假设，题目与相应的因子维度（构念）之间

有着明确的对应关系，所以小题单独计分对于解释某个构念应该更有意义（吴瑞林 2013：115）。著名统计学家 George Box（1979：236）曾指出："所有模型本质上都不准确，只是有些有用而已（Essentially, all models are wrong, but some are useful.）。"

高考英语阅读测试把阅读能力分解为五项测量目标（即五个维度），并依此编制了各类试题用以对其进行测量。本研究将从考生作答阅读测试样题的原始数据中提取对考生阅读能力有诊断作用的信息（即具体阅读技能的掌握情况），其预设是英语阅读能力（构念）是多维的。所以在运用认知诊断融合模型进行参数估计之前，研究者计划对本研究环境下的阅读能力维度进行分析。

2.2 认知诊断模型诞生的背景

Mislevy（1993：19-39）认为教育或心理测量理论大致可以分为两个阶段：标准测量理论（standard test theory）阶段和新一代测量理论（test theory for a new generation of tests）阶段。按照 Mislevy 的划分，经典测量理论（Classical Test Theory，CTT），甚至项目反应理论（Item Response Theory，IRT）都是标准测量理论阶段，属于能力水平范式（ability level paradigm）；而新一代测试理论则强调个体心理加工过程，也被称为认知水平范式（cognition level paradigm）（刘声涛等 2006；涂冬波等 2012：1；王卓然等 2012）。一般认为，《新一代测量理论》（*Test Theory for a New Generation of Tests*）（Frederiksen et al. 1993）的出版标志着认知水平范式理论的诞生。认知水平范式理论中尤以认知诊断模型（Cognitive Diagnosis Models，CDMs）为代表。

经典测量理论

人们通常认为 20 世纪 50 年代，以真分数理论（True Score Theory）为标志的经典测量理论（CTT）发展成熟。Gulliksen（1950）在《心智测验理论》（*Theory of Mental Tests*）一书中第一次用公理化的方式系统总结了标准化考试的原理和方法。CTT 的理论体系由三部分组成，真分数理论、测量信度理论与效度理论，其方法论体系包

括试题质量分析技术和测试命制的标准化技术两大部分（杨志明和张雷 2003：12）。CTT 主要以一整套测试试卷的视角来解释分数的含义，也就是以小题得分相加后的总分作为代表，单独一道小题的得分，不具有（对心理特质）解释的意义（余民宁 2009：7）。

CTT 范式下的试题质量指标（如难度、区分度、信度系数）与考生样本（能力水平）相互依赖。试题质量指标会因参加测试的考生群体样本不同而改变，考生能力水平的高低直接影响试题指标参数。同样地，考生的能力指标也严重依赖试题样本，试题的难易程度也直接影响考生能力指标参数。因此，同一份试卷受不同考生样本作答模式的影响，很难获得稳定的质量指标。尽管 CTT 存在诸多“缺陷”，但计算 CTT 的相关指标，并不要求大样本量，而且 CTT 使用的数学模型相对简单、易懂，大多数的常见的统计软件（如 SPSS、SAS、STATA 等）都采用 CTT 视角编写（Cohen et al. 2013：281；Green 2013：xiii）。Brown（2012：324）在回顾 CTT 在语言测试学科的应用时，认为 CTT 在当代语言测试的各个时期，如心理测量 / 结构主义（psychometric/structuralist testing）时期、综合性测试（integrative testing）时期以及交际语言测试（communicative testing）时期（Spolsky 1978；Hinofotis 1981；Brown 2005；Weir 2013），都扮演了“主导”的角色。尽管一些考试机构（如 ETS、Cambridge Assessment）都在积极运用 IRT 方法做质量分析以及试卷等值研究，但仍然会使用 CTT 对试卷（试题）质量进行初步分析，例如计算 CTT 范式下的相关参数（难度、区分度、信度系数、标准误差等）。Sawaki（2014：1160）也持相似观点，认为汇报 CTT 范式下的试卷质量指标仍然非常重要。即便在整体效度观时代，运用 CTT 对测验数据进行初步分析所得到的相关统计指标，仍然能为效度论证的不同阶段提供重要证据。

不得不承认，在教育与心理测量（包括语言测试）理论研究层面，概化理论（Generalizability Theory，GT）、项目反应理论，以及各类认知诊断模型早已迎头赶上。然而在现实世界的实际应用领

域，由于受到各种条件的制约，CTT 仍具其独特价值。从理论发展的角度来看，不具备 CTT 的理论常识，也就很难理解、掌握 IRT 和 CDMs 的理论要义及应用。部分研究（Lee et al. 2012；罗照盛 2012：20；Taylor 2013）表明，当考生样本足够大且能够较好地代表总体时，CTT 的参数估计结果也比较稳定，与 IRT、甚至 CDMs 对试题质量的估计结果趋同。

项目反应理论

为了克服 CTT 理论的局限，项目反应理论（IRT）应运而生。一般认为 ETS 的著名学者 Lord（1980）写的《应用项目反应理论解决实际测验问题》（*Applications of Item Response Theory to Practical Testing Problems*）一书的出版，正式宣告了项目反应理论的诞生（余民宁 2009：16；张厚粲和龚耀先 2009：122）。项目，即试题，“项目反应理论”是我国大陆心理测量界习惯的译法（漆书青和戴海崎 1992；罗照盛 2012），我国台湾地区心理测量界将其直译成“试题反应理论”（余民宁 2009），本书沿用了大陆学界已经习惯的译法，项目反应理论。IRT 是一种用于估测考生潜在特质（latent traits）的理论，其主要目的是把能够观察到的考生表现定位在无法观察到的潜在特质这一连续体上（Yen et al. 2006；Reise et al. 2013）。CTT 重点关注的是测验的总分（以及测验整体的质量水平），而 IRT 主要以试题的视角来解释测验分数的含义。应用 IRT，一般要符合两条主要的理论假设：潜在特质单维性假设（unidimensionality）以及试题局部独立性（local independence）假设。然而，现实世界中几乎所有真实考试数据都会不同程度地“违反”IRT 的基本假设（Ockey 2012：346）。Ellis & Ross（2014：1270-1271）提出了三种变通方法：一是使用更高级的 IRT 模型（如多维 IRT 模型克服单维性局限，或题组反应模型克服局部独立性局限）；二是对现有理论模型进行修订；三是“忽视”违反假设的情况，看最终模型拟合的结果是否支持某一种主要维度（特质）。

IRT 中的能力参数具有不变性特点，因为测试中的各小题对测验的贡献都是独立的。此外，还有项目参数（难度、区分度）估计不变