目　录

第一章
绪论

《中国英语能力等级量表》（以下简称《量表》）是首个面向我国英语学习者的英语能力测评标准，有助于为英语能力诊断提供共同标尺，有助于加强教学与考试之间的联动，通过综合改革实现二者的协同增效，实现我国外语教学提质增效。自 2018 年颁布以来，《量表》已逐步运用于语言测评、教学和学习等诸多方面（刘建达、杨满珍，2021），但相关应用较零散。本书尝试提出一个理论模型，以阐释《量表》如何在地方语境中指导英语教、学、评的一体化建设。具体而言，本书依据《量表》，结合认知诊断与标准设定方法，构建英语认知诊断测评反馈体系。此体系旨在发掘《量表》为学生提供个性化定性反馈报告的功能，从而促进英语教、学、评的有机融合，真正实现“车同轨、量同衡”。

1.1 研究背景

1.1.1 认知诊断测评与《量表》

相对于传统的常模参照性测试与准则参照性测试而言，认知诊断测评不再局限于笼统地报告考生成绩，而是能够科学、精确地诊断考生细化的语言技能结构与内在知识状态（He et al., 2021; Huebner et al., 2018; Li & Suen, 2013; Min et al., 2022; Wang & Qiu, 2019; 杜文博、马晓梅，2018；闵尚超、熊笠地，2019），既有助于教师针对性地提供个性化教学，也有助于学生自主学习效果的提高。尽管认知诊断测评

的重要性得到了专家学者的广泛认可（Alderson, 2005, 2010; Harding et al., 2015; Jang, 2010; Jang et al., 2013; Kunnan & Jang, 2009; Shohamy, 1992），也有一定的实证研究基础（Aryadoust, 2021; Li et al., 2016; Yi, 2017），但是认知诊断测评的长足发展还存在以下两个瓶颈问题。

（1）构念界定问题。认知诊断测评的理论基础为构念界定，即确定要测量的认知属性。长期以来，语言测试领域绝大部分测试设计均基于经典的交际语言能力模型（Bachman & Palmer, 2010）。但该模型集中于阐释语言交际能力，对语言理解能力的阐释不够细致，不能挖掘不同级别学习者完成听力或阅读任务时的认知过程，因此并不完全适用于认知诊断测评（Kim, 2015; 范婷婷、曾用强，2016）。相关理论框架的缺失导致真正意义上认知诊断测评的匮乏（Lee & Sawaki, 2009a），同时，也在一定程度上解释了为什么目前大部分研究仅限于将认知诊断模型应用到现有考试中，提取有限的诊断信息（Aryadoust, 2021; 闵尚超、熊笠地，2019）。

（2）诊断报告问题。认知诊断测评的最终目的是通过诊断报告，反馈教师与学生的情况，促进个性化教学。诊断报告承载着测试结果的描述及解释（Lee, 2015），是促进自主学习的内部催化剂（Butler & Winne, 1995），但相关研究明显不足（Lee, 2015）。部分学者做了相关尝试（Doe, 2015; Jang, 2009; Kim, 2015），探究如何呈现个体和群体反馈报告，但其研究结果基本止于对考生属性掌握模式的报道，无法将抽象的定量数字信息转换为更清晰、更丰富的定性反馈信息。相对于抽象的数字而言，图文并茂等定量定性信息结合的成绩报告往往更能被用户理解与接受（Tannenbaum, 2019）。

《量表》能在一定程度上解决认知诊断测评的这两个瓶颈问题。首先，相对于国际上其他能力量表而言，我国《量表》的一大突破为强调语言学习过程中认知能力的发展。《量表》在经典的交际语言能力模型（Bachman & Palmer, 2010）基础上，结合 Anderson & Krathwohl（2001）对 Bloom 教育目标分类学（修订版）中的认知和知识框架，用不同的认知行为表示能力等级的高低或能完成的语言交际活动的难度（He &

Chen, 2017; 刘建达、韩宝成，2018）。通过对“理解和表达意义”背后的各种“典型认知行为”进行描述，《量表》为认知诊断语言测试中的认知能力构念定义提供了原型。

其次，《量表》描述语包括三个成分：“行为”（performance）、“标准”（criteria）和“条件”（condition）。前两个成分缺一不可，第三个成分可空缺。听力和阅读描述语通过“行为”，即能做什么，体现学生的认知能力；通过“标准”，即输入或输出语言的特征，如话题的熟悉度、语法复杂度、词汇密度等，体现交际任务中具体的语言特征。通过将考试与《量表》建立关联，诊断报告结果将不再局限于数字化的掌握模式，而是可以为学生提供更清晰的个性化描述语，明确他们在听或读特定材料时能具体完成哪些认知任务。

因此，采用《量表》辅助认知诊断测评的构念界定与诊断报告的设计非常有必要，但在采用《量表》界定构念前，需更好地理解听力与阅读理解的过程与本质，厘清听力和阅读模态下各微技能间的异同。在采用《量表》指导诊断报告的设计前，需确保诊断测评的可靠性，挖掘适合用于诊断听力和阅读理解能力的模型。简言之，需考虑三个关键问题：第一，各认知属性之间的关系及其中间作用机理如何？第二，听力与阅读认知属性之间是否存在差异？第三，什么模型适合用于认知诊断语言测试？厘清这些问题有助于更好地将《量表》应用到诊断测评中。

1.1.2 考试对接《量表》

将考试与语言能力标准建立关联，有助于考试提供更清晰、丰富的反馈报告（Papageorgiou & Tannenbaum, 2016; Papageorgiou et al., 2019），有助于提升考试使用的效度（Dunlea, 2015）。这是构建个性化诊断反馈体系的核心环节。近十几年来，将考试与外部语言能力标准或能力量表进行对接的研究日益受到关注（Dunlea et al., 2019; Fleckenstein et al., 2020; Harsch & Hartig, 2015）。自《量表》发布后，如何将现有考试与其对接获得了国内外机构和研究人员的关注。目前，雅思、普思、托福等英美国家大规模英语考试与《量表》的对接结果已公布。

国内外相关文献中，对接研究亦不罕见，主要探究各类考试与《欧洲语言共同参考框架》（以下简称《欧框》）和《量表》的对接，探讨对接步骤的合理性和对接结果的确定（Dunlea et al., 2019; O'Sullivan et al., 2020; Papageorgiou et al., 2019; 揭薇，2019）。

（1）对接步骤。考试与标准对接包括四个阶段：框架熟悉（familiarization）、内容检视（specification）、标准设定（standard setting）和效度验证（validation）（Martyniuk, 2010）。以往对接研究发现一些问题，如：内容检视阶段中，《欧框》太过笼统，并未包括目标考试所考查的某些具体特征（Papageorgiou, 2010; Wu & Wu, 2010）；不同标准设定方法得出的对接结果不一致（Green, 2018; Kaftandjieva, 2010）；对接过程中较难判定构念无关因素对结果的影响（Papageorgiou, 2010），因此对接研究的一个重要问题就是探讨对接结果是否具有准确性和可靠性。

（2）对接结果。国际上比较重要的外语水平考试向来都重视与各语言能力标准的对接，关注测试对接结果的报道。早在 2008 年，美国的托福考试就率先完成了与《欧框》的对接，以帮助测试者和决策者根据相应的能力量表等级对测试分数进行更全面的解释（Tannenbaum & Wylie, 2008），随后又于 2014 年根据使用者反馈修订《欧框》各等级的最低分数线，以保障分数线的合理性（Papageorgiou et al., 2015）。英国的雅思、普思和培生学术英语考试通过与语言能力量表的对接为其考试效度提供了证据，也使其成绩报告能包含考生在各语言技能上的得分及所达到的《欧框》等级，更真实详尽地描述考生在不同层面的语言水平（O'Sullivan, 2015; Taylor, 2004）。作为移民大国，加拿大非常重视语言水平，完成了各项法语考试（如法语水平考试 TEF、法语学习证书 DELF 等）与《欧框》及《加拿大语言能力标准》之间的对接（Casanova & Crendal, 2011; North & Piccardo, 2018）。

目前相关研究主要集中探讨考试对接量表步骤的合理性、分数段与等级划分，鲜有研究探讨如何根据量表描述语为各分数段的考生提供定性反馈，让学生、家长以及用人单位清楚了解该分数段或该等级

的学生具体能做什么。因此，有必要采用《量表》设计定性反馈报告，提高报告的可读性和可理解性。但是在设计报告前，需解决的核心问题是对接本身是否具有效度，尤其是对接结果是否具有一致性。这包括两个关键问题：第一，采用同一标准设定方法进行对接时，标准设定专家自身和之间的一致性如何？第二，采用不同标准设定方法进行对接时，不同标准设定方法得出的结果之间的一致性如何？

1.1.3 个性化反馈与《量表》

诊断测评的核心是通过为学生提供个性化的反馈报告，实现后续针对性教学。认知诊断反馈报告已在大规模教育测评中得到应用（Bradshaw & Levy, 2019）。在语言测试领域，学者们对通过认知诊断测评来提供诊断反馈的兴趣与日俱增（Javidanmehr & Sarab, 2019; Kim, 2011; Li et al., 2016; Mirzaei et al., 2020; Xie, 2017; Yi, 2017）。例如，美国教育考试服务中心（以下简称 ETS）的研究人员使用认知诊断测评为每位考生提供托福考试听力和阅读部分（Sawaki et al., 2009b）的个性化反馈。此外，Zhang et al.（2019）尝试结合认知诊断和尺度锚定的方法来优化雅思考试阅读部分的分数报告。

但是，目前的认知诊断反馈报告通常仅限于描述考生个人的认知表现，很少关注考生可以理解的输入材料的特征。根据 Harding et al.（2015）和 Alderson（2005）的观点，即使是低水平的学生也能正确回答测量高阶认知技能（如推理和归纳大意等）的题目，即使是高水平的学生也会错误回答测量低阶认知技能（如理解词意等）的题目。决定学生题目表现的因素不仅包括认知技能的难度，还有文本的语言特征（Alderson, 2007）。因此，若能向分数使用者同时提供有关任务特征的描述信息和考生在认知技能上的表现，则能更好地了解考生的语言能力。同时，仅有高质量的反馈报告未必促学，除非学生能理解和使用反馈报告，并在此基础上采取实际行动，而且老师能基于诊断结果给予后续针对性更强的教学活动（Ajjawi & Boud, 2017; Gibbs & Simpson, 2004; Vogt et al., 2020; Winstone et al., 2019）。简言之，诊断测评研究有待进一步深入探讨的两个关键问题是个性化反馈报告的研

制以及基于个性化反馈的后续针对性教学和学习效果。《量表》的使用有助于更好地回答这两个关键问题。

（1）个性化反馈报告的研制。结合认知诊断与标准设定方法能为考生提供高质量个性化反馈报告。此类反馈报告既能反映出理解特定级别的书面和口头文本的能力（Green, 2018; Powers et al., 2017），又能反映出认知加工的强弱项（Jang et al., 2015; Kim, 2015）。但据笔者所知，目前鲜有研究将这两种方法结合起来，为考生提供个性化反馈，以促进后续的针对性教学和学习。这可能是因为学者尚未厘清基于标准设定的考生水平分类和基于认知诊断测评的微技能掌握情况分类之间的关系。由于这两种方法用的评分与分类方法不同，可能会出现以下情况：某些被归入高水平的考生有可能被诊断为并未掌握所有微技能，而一些被归入低水平的考生却被诊断为掌握了所有微技能（Liu et al., 2018）。这种情况下，反馈报告提供者很难向反馈报告使用者解释这两种信息的不一致。但是，使用者同时需要这两种反馈信息，来为后续针对性教学做出决策，并准备教学材料（Jang et al., 2019）。例如，Hyatt & Brooks（2009）对英国大学的考试利益相关者进行了采访，发现 74% 的受访者认为被录取的英语学习者入校后需要额外的语言支持，但 64% 的受访者表示雅思成绩报告没有提供此类诊断信息。因此，研究者有必要关注标准设定和认知诊断测评方法的一致性，从而探讨结合这两种方法以提供个性化反馈的可行性。

（2）个性化反馈报告的使用效果。尽管学生作为反馈报告的潜在主动接收者，有责任积极回应反馈报告（Winstone et al., 2017），但是他们对反馈报告的使用会受多个因素的影响。反馈报告素养不足可能导致学生无法有效使用反馈报告，尤其当反馈报告包含复杂的术语和抽象的图表时（Carless & Boud, 2018; Underwood et al., 2010; Zapata-Rivera et al., 2016）。更糟糕的是，由于教师和学生之间的权力关系差异，学生可能无法表达对反馈报告的异议和误解（Leighton, 2019）。一种可行的解决方案是将诊断结果与后续的针对性教学联系起来，但

是在实践中此类探索寥寥可数（Elder & Read, 2015a, 2015b; Oral English Proficiency Program, 2015, 2019）。

虽然以往研究者设计了不同形式的针对性教学活动，旨在加强诊断测评结果在教学方面的作用并带来积极后效（Lee, 2015; Mason & Singh, 2010），但是这些活动的设计缺乏统一的语言能力标准作为参考。就教学而言，语言能力标准（如《量表》）可以帮助教师制定更为明确、详细、切实的教学目标，帮助教师选择合适的教学内容和教学方法。就学习而言，学生可以根据《量表》描述语，结合自己的实际情况制定更为明确的学习目标，选择适合的学习材料，提升自主学习能力。尽管目前有少量研究探究针对性教学，但鲜有研究探索针对性教学活动的有效性，更少有研究探究基于语言能力标准的个性化反馈报告与基于语言能力标准的针对性教学活动结合起来的有效性（Min et al., 2022）。

因此，基于《量表》的个性化反馈及其后续针对性教学的有效性是一个值得探究的话题，包括两个关键问题：第一，如何设置基于《量表》的个性化反馈报告？第二，基于《量表》的个性化反馈报告及其后续针对性教学的效果如何？

1.2 研究目标与意义

本研究的总体目标是：基于《量表》，以校本考试为依托，结合认知诊断与标准设定方法，构建英语认知诊断测评与个性化反馈报告体系，并提出基于《量表》的诊断反馈促教促学模型，以期指导《量表》在教、学、评中的长期应用。基于总体目标，本研究探究七个具体分项目标，该七大分目标分别围绕认知诊断测评、考试对接《量表》、基于《量表》的个性化反馈报告三大主要模块展开。

• 认知诊断测评

第一，互补性机制。从认知属性层面探究听力理解的过程与本质，重点着眼于采用互补型与非互补型认知诊断模型，研究听力理解过程是否存在互补性机制及其具体体现，以期阐释各认知属性之间的关系及其中间作用机理。

第二，跨模态稳定性。采用认知诊断模型，重点探索听力与阅读理解过程中共有的认知属性的表现稳定性（相似性）和可变性（差异性），挖掘听力和阅读模态下各微技能间的异同，以期建立更明晰的听力和阅读技能机制。

第三，模型选择。采用不同类型的模型——认知诊断模型与多维项目反应理论模型，拟合听力测试数据，以期找到适用于捕捉语言能力的多维性和连续性的最佳模型类型。

• 考试对接《量表》

第四，同一标准设定方法的一致性。采用改良 Angoff 为标准设定方法，收集标准设定和实际考试数据，从同一标准设定方法的内部一致性以及反馈是否能提高一致性的角度，探讨某校本考试阅读卷与语言标准对接的效度问题，以期为后续结合标准设定和认知诊断测评方法开发的个性化反馈报告提供效度证据。

第五，不同标准设定方法的一致性。采用改良 Angoff 和对照组法为标准设定方法，收集标准设定和教师评判数据，从不同标准设定方法间的一致性，探讨某校本考试听力卷与语言标准对接的效度问题，以期为后续结合标准设定和认知诊断测评方法开发的个性化反馈报告提供效度证据。

• 基于《量表》的个性化反馈报告

第六，基于《量表》的个性化反馈报告研发。从认知诊断和标准设定结果的一致性角度，探究结合认知诊断与标准设定方法，制定基于《量表》的个性化诊断反馈报告的可行性，以期解决大规模高风险考试无法提供个性化反馈的困境。

第七，基于《量表》的个性化反馈报告的使用效果。开展为期近五个月的历时研究，检验使用基于《量表》的个性化反馈报告和基于《量表》的针对性教学干预是否有助于学生语言能力的提升，以期为本书提出的基于《量表》的诊断反馈促教促学模型提供实证依据。

1.3 章节构成

本书第一章为绪论，简要概括本书的研究背景、研究目标与意义。第二章介绍《量表》听力与阅读理解能力分量表的理论模型、描述语框架、各级别描述语典型特征，以及相对于国外语言量表和标准的几大突破。第三章阐述听力与阅读理解能力的本质与异同。第四章概述认知诊断测评的起源与基础、模型、操作步骤，以及其在语言测评领域中的应用、发展与挑战。第五章介绍考试对接标准的起源与基础、操作步骤，以及发展与挑战。第六章详细介绍标准设定方法的基本步骤、方法，以及发展与挑战。第七章提出本研究构建的基于《量表》的诊断反馈促教促学模型。第八至十四章呈现七个实证研究，分别探究认知诊断测评中听力能力的互补性机制（第八章），听力与阅读微技能跨模态稳定性（第九章），认知诊断模型选择（第十章），考试对接《量表》的效度，包括同一方法的内部一致性（第十一章），不同方法间的一致性（第十二章），基于《量表》的个性化反馈报告研发（第十三章）以及使用效果研究（第十四章）。第十五章总结本书的主要发现、理论价值与实践意义，并指出未来研究方向。

1.4 小结

本章首先简要介绍了《量表》的概况，然后通过梳理认知诊断语言测试、考试对接标准、个性化反馈等方面的研究现状与不足，指出《量表》有助于解决以上领域的瓶颈问题，且能在诊断、反馈、干预中发挥桥梁作用，促进英语教、学、评的有机融合。接着基于研究现状，提出本研究的总目标和七个分目标，以及各自的研究意义。最后概述本书的章节构成。