探索语言水平测试的认知诊断改造和深度分析：以广州市英语学业考试为例*

林燕婷[1] 陈慧麟[2] 陈劲松**[3]
（[1]中山大学心理学系，广州，510006）（[2]上海外国语大学国际教育学院，上海，200083）
（[3]中山大学心理学系，广州，510006）

摘 要 本研究探索在通用认知诊断模型和相关检验方法的基础上对现有语言水平测试进行诊断改造和分析，分三步进行探索：（1）探索对语言水平测试不同的属性和Q矩阵构建途径；（2）探索对语言水平测试基于通用模型的建模和效度验证；（3）探索对语言水平测试建模后续的深入分析。研究发现：属性分布和总分分布划分的学生水平一致性较高；学生对属性掌握存在性别差异且属性间的难易层级不同；属性模式分布进一步验证了语言属性间关联程度较高以及通用认知诊断模型和相关检验方法对语言测试的适用性。三步式的建模分析可作为对语言水平测试进行认知诊断改造的参考。

关键词 认知诊断模型 英语成就测试Q矩阵 G–DINA 拟合指数

1 引言

认知诊断模型（cognitive diagnosis model, CDM）和以其为基础的认知诊断评估（cognitive diagnostic assessment, CDA）在教育和心理测量领域受到越来越多的关注。相对于强调群体中个体之间比较的传统测量理论，CDA更关注细颗粒度的能力和知识结构，并可以绝对的标准判断个体的认知状态（Rupp, Templin, & Henson, 2010）。CDM中认知属性是潜分类变量，其分类定义来自于外部，并通过Q矩阵(Tatsuoka, 1983)引入模型，诊断个体确切能力的掌握情况，测量学生在目标领域的学习状态，据此可以提供有效的反馈，帮助学生改进学习或教师改进教学。

本研究的目的是结合认知诊断理论和方法研究的进展，探索对语言水平测试的认知诊断改造和深度分析。英语学业考试是属于能力水平考试和学业成就考试相结合的考试，其目的是为了考查学生的英语能力，属于效标参照测试，对其进行认知诊断改造分析更能够体现测试的目的。本研究探索在通用认知诊断模型和相关检验方法的基础上对现有语言测试进行三步式的建模分析：（1）探索对语言水平测试不同的属性和Q矩阵构建途径，结合英语语言研究专家、英语课程标准以及考试大纲所规定的初中阶段英语毕业水平的能力要求，对初中生英语测试进行分析，并根据分析所得的认知属性集构建Q矩阵；（2）探索对语言水平测试基于通用模型的建模和效度验证，采用通用的认知诊断模型和相关检验方法按照推荐步骤（图1）进行建模并结合专家意见修订Q矩阵；（3）探索对语言水平测试建模后续的深入分析，包括水平划分分析、平均掌握情况及性别差异的评估、属性难易层级分析和属性模式分布等。在科研上，本研究通过真实数据进一步探索和完善对现有测试进行认知诊断改造的方法，并探索如何进行深度的分析；在教学上，本研究希望能够更准确地诊断学生的英语学习状态，从而有利于帮助学生改进学习或帮助教师改进教学。

2 文献背景

2.1 认知诊断模型及建模检验的发展

CDM包含饱和模型和简约模型。饱和模型有G-DINA（generalized deterministic input, noisy “and” gate）、LCDM和对数CDM (de la Torre, 2011)，均包容

* 本研究得到中山大学国家高等教育质量常态监测数据中心（高等教育研究院）开放基金的资助。
** 通讯作者：陈劲松。E–mail：jinsong.chen@live.com
DOI:10.16719/j.cnki.1671–6981.20180434

所有简约模型。简约模型有很多种，可以粗略分为补偿型（如 LLM 和 A-CDM）和联结型（如 DINA）。本研究采用G-DINA模型(de la Torre, 2011),其公式如下:

$$P(\alpha_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^{*}} \delta_{jk}\alpha^{*}{}_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \delta_{jkk'}\alpha^{*}{}_{lk}\alpha^{*}{}_{lk'} + \cdots + \delta_{j12\ldots K^{*}j}\prod_{k=1}^{K_j^{*}} \alpha^{*}{}_{lK^{*}j}$$

，其中 j 代表题目，k 代表题目所测认知属性的数量，$P(\alpha_{lj}^{*})$代表特定属性的主效应以及他们间交互作用的总和，δ_{j0}是猜测效应，δ_{jk} 指掌握一个属性α_{lk}时的主效应，$\delta_{jkk'}$表示属性α_{lk}和$\alpha_{lk'}$的交互作用效应，$\delta_{j12\ldots K^{*}j}$表示属性$\alpha_1 \ldots \alpha_{k^{*}j}$的交互效应。该模型题目参数的数量达到最大化，覆盖了认知属性所有可能的关联情况，学生掌握必需属性的数量越多，回答正确率就越高，由每个必需属性的主效应以及属性间的交互作用构成的，适用于多元抽象的语言诊断测试(Chen & Chen, 2016a, 2016b)。

随着各种认知诊断模型的不断开发，认知诊断评估数据与模型的拟合检验也在不断的发展完善。在采用 CDM 进行诊断的一个普遍且关键的步骤是，需要该领域的专家构建 Q 矩阵以确定题目和属性之间的关系 (Tatsuoka, 1983)。通过 Q 矩阵，实质性的属性知识能够集成整合到建模过程中，从而提供了丰富的诊断信息。因此，CDM 的模型评价包括判断模型、Q 矩阵与数据的拟合程度，其检验方法可分为整体水平的拟合指标和局部水平的拟合指标。整体水平的拟合检验是基于标准残差的绝对拟合检验，包括 Fisher 转换观察和预测的项目对相关度标准残差 *zr* 和项目的对数优势比标准残差 *zl*，该指标能在整体水平上检验模型和 Q 矩阵的有效性，从而判断该测试以及其所产生的诊断信息是否有效。局部水平的拟合指标主要有：属性水平的标准残差的均方根值（*sr* 或 *sl*）；题目水平标准残差均方根值（sr_j 或 sl_j）(详见：Chen, 2017; Chen & de la Torre, 2013)。此外，针对属性分类假设的检验使用 MAD 指标，该指标大于 .35 时属性偏离分类假设较为严重，可能不适合 CDM（Chen, de la Torre, & Fu, 2015）。

基于以上整体水平和局部水平的检验方法，本研究对现存的检验步骤进行了修改，提出了如图 1 所示的检验步骤，能够从测试的整体和局部综合检验 Q 矩阵并进行修订：首先要确定 Q 矩阵的完整性及可识别性，即一定数量的学生数以及测试题目；接着检验属性分类假设 MAD 值，再按照 Chen (2017) 的 Q 矩阵的四步法检验：基于 *zr* 和 *zl* 判断 Q 矩阵是否在整体水平上可接受，如果拒绝（基于 .01 的显著水平），进一步评估属性水平是否存在误设（*sr* 或 *sl* >1.5，即存在误设），不存在时根据题目水平标准残差的均方根值中（sr_j 或 sl_j）的极大值（*mr* 或 *ml*）所推荐的题目，在指标 Δsr 或 Δsl 以及专家意见的协助下修订该题目所考查的属性，再检查 Q 矩阵在整体水平上是否可接受。若修订题目超过五题，基于整体水平的 *zr* 和 *zl* 依旧不能接受，其属性定义可能存在问题或存在大量误设的题目，就需要进行整体水平的修改（如重新定义属性、重新建立 Q 矩阵）。

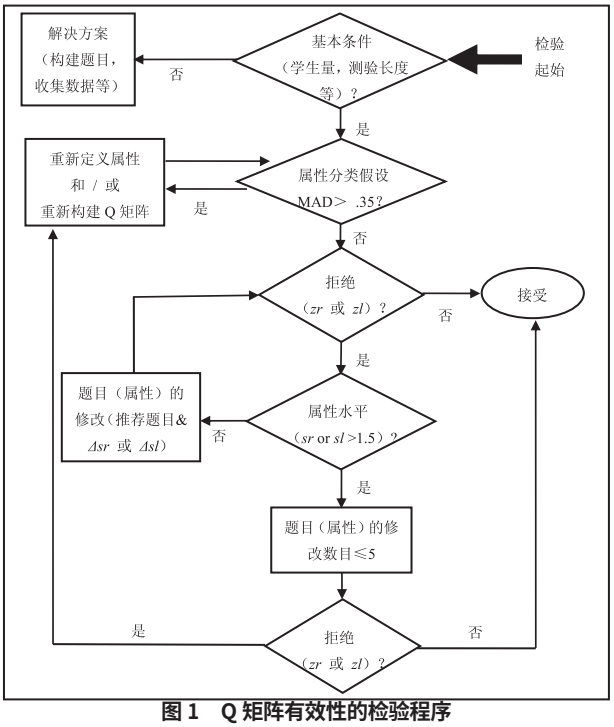

图 1　Q 矩阵有效性的检验程序

2.2 语言测试的认知诊断改造

目前，已有部分研究对英语测试进行认知诊断改造，但大部分只是是英语阅读题型，且针对国内被试群体的研究很少。针对国外被试群体的研究有：采用 GDM 模型 (von Davier, 2005)，融合模型 (Jang, 2009; Kim, 2015)，LLM(陈慧麟，赵冠芳，2013)，G-DINA 模型 (Chen & Chen, 2016a, 2016b) 对英语阅读测试进行认知诊断。以上普遍发现存在多个相对独立又相互关联的阅读属性，属性间具有一定的层级关系，但会相互影响，并且不一定适合较为简单的补偿型或联结型模型，

这与语言技能的多样性、综合性和相关测试的复杂性相一致（Heaton, 1988）。但现存的研究检验没有对局部（尤其题目水平）进行检验和修订，也没有检查属性的分类假设是否成立且没有特定的步骤和策略；研究只以阅读题型考查的能力为主，其他语言技能极少牵涉且没有做进一步的诊断分析。

3 研究方法

英语学业考试覆盖的内容较多，本文选取有一定研究基础的两部分内容进行探索：语言知识与运用和阅读理解。考虑到当前 CDM 的理论研究目标属性个数一般低于 10 个，而且语言知识与运用和阅读理解这两部分内容相对独立，研究将分别对两部分内容进行建模，再合并分析。

3.1 被试与测试材料

本研究抽取了参加 2015 年广州市初中毕业生英语学业考试的2718名学生的作答数据作为样本，其中，男生 1388 名，女生 1330 名。测试材料为该年试卷中的语言知识与运用和阅读理解题型，一共 40 道题目。

3.2 初步的认知诊断改造

3.2.1 认知属性的确定

结合英语语言学专家、前人的属性定义研究 (陈晓扣 , 李绍山 , 2006; Alderson & Lukmani, 1989; Chapelle, 1999; Larsen-Freeman, 2003; Munby, 1981)、《义务教育英语课程标准（2011 年版）》（教育部，2012）以及《广州市初中毕业生学业考试英语考试指导书》（广州市教育局教学研究室，2011）对 40 道题目进行分析，分别确定了考试大纲法以及专家确定法的两个认知属性集。

考试大纲法的认知属性是英语语言专家结合《义务教育英语课程标准（2011 年版）》和《广州市初中毕业生学业考试英语考试指导书》所规定的初中阶段英语毕业水平的能力要求，对学业考试中的 40 道题目进行分析，得出的与词汇、语法和阅读技能有关的 11 个认知属性（A1-A11）。专家确定法是英语语言专家根据考试题目、前人已研究的属性定义，进行分析讨论构建的全新的属性集，包括 4 种属性类别，即词法、句法、完型和阅读，共 13 个认知属性（表 1）。

3.2.2 测试 Q 矩阵的构建与界定

分别根据考试大纲法和专家确定法的属性定义，结合具体的题目内容，邀请英语专家构建 Q 矩阵。Q11 和 Q12 是根据考试大纲法的认知属性和题目内容所构建的 Q 矩阵。试题分为语言知识与运用和阅读理解这两部分，由于语言知识与运用主要考察词汇和语法，Q11 中只涉及到词汇和语法属性，共有 6 个属性；阅读理解主要考察了阅读技能，因此 Q12 中只涉及了 5 个阅读属性。Q21 和 Q22 是根据语言专家确定的认知属性定义，结合题目内容，分别对语言知识与运用和阅读理解构建 Q 矩阵。由于属性测量的次数不能过少，删除了在本测试中只考查一次的属性 A11 和 AR5 以及第 50 题。接着，邀请了 4 位初高中英语教师根据教学经验分别对以上 Q 矩阵的每个题目所考查的属性进行同意或不同意的评价（不同意需要注明原因）。矩阵 Q11 和 Q12 的同意率的区间在 25%-100%，均数分别为 80% 和 82.5%；矩阵 Q21 和 Q22 的同意率区间均为 25%~100%，均数分别为 83.75% 和 81.25%。教师的评价结果总体上表明了初始 Q 矩阵具有初步的有效性。

表 1 专家确定法的认知属性

属性类别		属性名称	属性定义
AW1	词法	正确搭配词汇	识别并合成搭配正确的词汇或短语。
AW2	词法	辨析形近词汇	辨析形式上相近的单词或短语在意义上的差异。
AW3	词法	辨析近义词汇	辨析意义上相近的单词或短语间存在的细微差别。
AW4	词法	正确判断词性	根据语法规则和意义确定单词的词性
AS1	句法	正确使用非谓语动词	在句子中适当的地方以适当的形式使用非谓语动词。
AS2	句法	分析动词时态和语态	识别动词时态和语态的正确形式，并正确地使用。
AS3	句法	分析句子内部的主从和并列关系	分析句子内部的主从和并列关系，并使用正确的连接词和语序。
AC1	完型	理解语篇中与空缺有关的非显性信息	对篇章中非直接的语义信息进行归纳，补出空缺处的信息。
AR1	阅读	提取语篇中的显性信息	从文中查找出与正确回答相同或相似的文字，并以此作答。
AR2	阅读	阐释或转述显性信息	对篇章中的文字进行解释或转述，理解这些文字所体现的概念和逻辑关系。
AR3	阅读	概括全文或段落大意	概括全文或段落的主旨大意，得出整体性的理解。
AR4	阅读	对隐性信息进行推理	推断出非文字体现的有关篇章内容的隐性信息。
AR5	阅读	对篇章进行评价	对篇章风格、意义、以及作者态度等非内容信息进行评价。

3.3 检验与分析

对测试Q矩阵的检验分析采用Ox软件(Doornik, 2003），并遵循图 1 的检验程序和专家意见对 Q 矩阵进行修订，以获得可接受的 Q 矩阵。最后基于可接受的 Q 矩阵，进行认知诊断和深入分析。

4 结果

4.1 建模检验

表 2 Q 矩阵的属性分类假设 MAD 值

	考试大纲		语言专家	
	Q11	Q12	Q21	Q22
MAD 值	.31	.34	.30	.34

采用饱和 G-DINA 模型 (de la Torre, 2011) 进行诊断，得到每个 Q 矩阵的属性分类假设 MAD 值均小于 .35，如表 2，可进一步的检验分析。

考试大纲属性集构建的 Q 矩阵的初始拟合结果和语言专家确定属性集所构建的 Q 矩阵的初始绝对拟合结果 *zr* 值均大于临界值（.01 显著性水平），不可接受。按照图 1，尝试根据 *sr* 或 *sl* 所推荐的修订题目并结合专家和教师小组的意见后修订 Q11 和 Q12（即先通过局部拟合指标找出最可能需要修订的题目，然后让专家小组给出题目修改意见，再通过整体拟合指标判断意见是否能改善拟合），但修订多题（分别多于 5 题）后其绝对拟合结果仍没有变好（仍远高于临界值），按照图 1 的程序，Q1 矩阵存在属性认定的问题，因此停止进一步的诊断分析。按照同样的方式尝试对 Q21 和 Q22 进行修订，分别修订若干题后 Q 矩阵的绝对拟合指标均可接受（表 3）。两个 Q 矩阵所修订题目数量分别为 4 和 3 题。试卷的信度 α 为 .921，题目的猜测概率平均为 .35，失误概率平均为 .11。为检验修订的题目属性情况是否合理，邀请了 5 位初高中英语老师（部分为新老师）对修订的每个题目所考查的属性分别进行是否同意的评价，同意率区间为 80%-100%，均数为 88%。以上结果表明对专家确定属性集构建的 Q 矩阵经过适当修订后可接受。

表 3 矩阵 Q21 和 Q22 的修订后的绝对拟合结果

	Q21		Q22	
	相关	优势比	相关	优势比
zr 或 *zl*	3.89	3.80	3.68	3.34
sr 或 *sl*	1.40	1.34	1.43	1.37

注：显著性水平 p=.01 的临界值：Q21：4.02；Q22：4.02

4.2 进一步分析

基于可接受的 Q21 和 Q22 矩阵，分别对其进行认知诊断改造的结果进行深入分析。

4.2.1 学生水平划分分析

根据认知诊断分析的结果，按照考生掌握的属性个数（12 个认知属性）和传统的测试总分（满分 55 分）将考生划分为三个等级并进行比较。优秀的学生（90%）掌握的属性个数大于等于 11 个或测试得分大于等于 50 分，及格的学生（60%）掌握属性个数大于等于 7 个或测试得分大于等于 33 分，结果如表 4。两种划分方法按照是否及格划分得到的学生水平一致性达到了 84.58%，其中，一类误差（属性划分及格而总分划分不及格）所占比例为 1.36%，二类误差（属性划分为不及格而总分划分为及格）为 14.06%，相比较于属性划分，总分划分会高估学生的能力；划分学生水平是否优秀的一致性达到 95.14%,一类误差(属性划分优秀而总分划分不优秀）占比为 3.61%，二类误差（属性划分不优秀而总分划分优秀）占比为 1.25%。这表明传统测试总分划分下的学生能力水平与认知诊断学生掌握属性情况的划分结果较为一致。

4.2.2 属性的平均掌握情况及性别差异

通过模型计算获得学生对各认知属性的平均掌握概率，如表 5。初中生对属性 AR3“概括全文或段落大意”的平均掌握概率最低，其余属性的平均

表 4 学生水平划分情况的比较

	属性		总分		一致性%	I 类误差	II 类误差
	人数	比例	人数	比例			
及格	1097	40.36%	1442	53.05%	84.58%	1.36%	14.06%
优秀	547	20.13%	483	17.77%	95.14%	3.61%	1.25%

注：I，II 类误差均以属性划分为真判断

表 5 属性的平均掌握概率

属性编号	概率	属性编号	概率	属性编号	概率
AW1	.47	AS1	.40	AR1	.51
AW2	.50	AS2	.63	AR2	.57
AW3	.66	AS3	.41	AR3	.38
AW4	.59	AC1	.67	AR4	.43

掌握概率均高于 .40。

进一步探讨认知属性的掌握情况是否存在性别差异。属性 AR4 掌握概率不存在显著的性别差异。其余属性均存在显著的性别差异，属性 AC1 女生的掌握概率显著低于男生的掌握概率，*ts*(1，2716)=2.17，*p*=.030，*d*=.09；其余属性女生的掌握概率显著优于男生的掌握概率，*ts*(1，2716)> -2.51，*ps*<.012，*ds* ≥ .11，如图 2 所示。

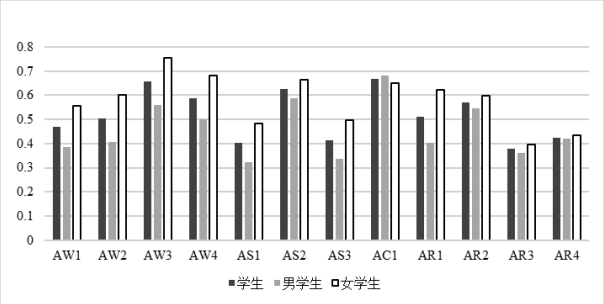

图 2　属性掌握概率情况及性别差异

4.2.3　属性间难易程度层级探讨

通过对每一个学生每一个认知属性的掌握情况进行两两比较，计算每属性之间掌握概率的优势比（Odd Ratio），探索属性的难易层级关系，如表 6。若两个属性（A/B）的优势比为 1，表明学生中掌握这两个属性的难易均势；若优势比为 2，学生中属性 A 的掌握情况优于属性 B 的掌握情况的人数是学生中属性 B 的掌握情况优于属性 A 的掌握情况人数的 2 倍，也就是说属性 A 较易掌握，以此类推。

词法中，属性 AW1、AW2 和 AW4 难易均势，而 AW3 较易。句法部分，属性 AS3 最易，属性 AS1 和 AS2 难易均势。学生对阅读属性的掌握情况出现两极，对属性 AR1 和 AR2 掌握难易均势，属性 AR3 和 AR4 的难度逐渐增加，前两个较易掌握。

4.2.4　属性模式的分布

诊断分析获得属性模式分布情况如下表 7。语言知识与运用部分一共有 8 个认知属性，其组成的潜质类型有 256 种，其中，大于 .01 的潜质分类如下表，认知状态“11111111”所占比重最大（29%），即全部掌握的类型所占比例最大，说明了语言技能互为关联。阅读理解的认知属性有 4 个，共有 16 种潜质类型，具体结果如下表。综合模式分布，发现多数考生要么同时掌握要么同时不掌握绝大多数属性。

4.3　外部效度检验

为检验本研究认知诊断的有效性，以学生的

表 6　属性难易层级比较

层级比较	优势比	层级比较	优势比	层级比较	优势比
AW3:AW4	3.03	AS3:AS1	2.06	AR1:AR2	1.05
AW4:AW2	1.57	AS1:AS2	1.04	AR2:AR3	3.26
AW2:AW1	1.33			AR3:AR4	1.83

表 7　属性模式的概率分布

认知状态	分布概率	认知状态	分布概率
"11111111"	.29	"0100"	.29
"00001001"	.16	"1000"	.25
"00000001"	.16	"1111"	.23
"00010010"	.06	"0011"	.12
"10011010"	.03	"0110"	.03
"10010010"	.03	"0010"	.02
"00010110"	.03	"0000"	.01
"10011110"	.02	"1101"	.01
"00111111"	.02	"1010"	.01
"10010110"	.02	总和	.99
总和	.84		

注：13 位数字从左至右分别是 AW1、AW2、AW3、AW4、AS1、AS2、AS3、AC1、AR1、AR2、AR3、AR4

表 8　认知诊断结果与总成绩相关

	能力	AW1	AW2	AW3	AW4	AS1	AS2	AS3	AC1	AR1	AR2	AR3	AR4
听力	.77**	.73**	.76**	.77**	.78**	.71**	.53**	.71**	-.07**	.77**	.15**	.27**	.16**
写作	.79**	.77**	.80**	.83**	.82**	.72**	.53**	.74**	-.13**	.79**	.16**	.23**	.12**

注：** 在 .01 水平上显著相关

听力和写作成绩作为外部效标，探讨诊断所得后验的属性掌握概率 (posterior probability of attribute mastery) 与效标的相关，如表 8，除了属性 AC1 与效标呈显著的负相关，其余均显著正相关，表明诊断结果具有较好的外部效度。

5 讨论

本研究探索在通用的认知诊断模型和相关检验方法的基础上对现有语言水平测试进行诊断改造和分析，分三步进行探索：（1）探索对语言水平测试不同的属性和 Q 矩阵构建途径；（2）探索对语言水平测试基于通用模型的建模和效度验证；（3）探索对语言水平测试建模后续的深入分析。建模检验发现根据考试大纲定义的属性集和 Q 矩阵在通用模型框架下与数据的拟合度较差，原因可能是考试大纲相对于实际考试内容更为宽泛。根据语言测试文献研究定义的属性集和 Q 矩阵其最终拟合结果可接受。因此，对现有测试进行认知诊断改造分析时应注意属性集的确定和 Q 矩阵的构建，将来研究可进一步探讨如何结合考试大纲和前人文献构建更合理的属性集和 Q 矩阵。总的来说，三步式的建模分析可作为对语言水平测试进行认知诊断改造的参考。

考虑到学业考试的高利害性和两种及格划分方法的一致性，不建议改变目前学业考试的计分方式，但建议在将来的考试报告中增加被试对每个属性掌握的概率，在整体的测试分析报告中也可以增加基于属性的分析，以提供更加细致和全面的诊断性信息，提高学业考试的效用。限于目前的 CDM 理论研究目标属性数量，本文将相对独立的测试内容单独建模，然后再合并分析。将来理论条件成熟时可合并所有内容建模分析，也许会得到更加精确的测量结果。另外，如审稿人提出，阅读部分题目可能存在局部依赖性，将来的应用研究可参考相关的方法文献(如詹沛达 , 李晓敏 , 王文中 , 边玉芳 , 王立君，2015；Hansen, Cai, Monroe, & Li，2016)。

采用 CDM 分析测试时，理想情况下希望 Q 矩阵满足完备性条件 (Köhn & Chiu, 2017)。两个 Q 矩阵 Q21 和 Q22 均符合通用模型的完备性要求。但对现存测试进行改造时，受限于现有条件比如属性定义和题目，满足完全完备性的 Q 矩阵不一定能建立。从另一角度而言，如果属性数量比较多而完备性的违反并不严重（比如只限于个别属性），只会影响到少量属性模式 (patterns) 的估算，从整体而言属性分布和相关的诊断信息仍将具有一定的价值。此外，研究表明在 CDM 中引入题目作答时间 (response time)，其模型参数的估算结果更精确，能够在一定程度上提高属性及属性掌握状态的判准率 (Zhan, Jiao, & Liao, 2017)，尤其随着计算机测试的盛行，这可能成为将来的趋势。

致谢：衷心感谢广州市教育研究院提供数据。

参考文献

陈慧麟 , 赵冠芳 . (2013). 认知诊断的应用——语言测试研究的新阶段 . *外语测验与教学 , 2,* 1–9.

陈晓扣 , 李绍山 . (2006). TEM–4 完型填空测试结构效度研究——答题过程分析法 . *现代外语 , 29(1),* 71–77.

广州市教育局教学研究室 . (2011). *广州市初中毕业生学业考试指导书 – 英语* . 广州 : 广东教育出版社 .

詹沛达 , 李晓敏 , 王文中 , 边玉芳 , 王立君 . (2015). 多维题组效应认知诊断模型 . *心理学报 , 47(5),* 689–701.

中华人民共和国教育部 . (2012). *义务教育英语课程标准 : 2011 年版* . 北京 : 北京师范大学出版社 .

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in A Foreign Language, 5*, 253–270.

Chapelle, C. A. (1999). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chen, H. L., & Chen, J. S. (2016a). Exploring reading comprehension skill relationships through the G–DINA model. *Educational Psychology, 36(6),* 1049–1064.

Chen, H. L., & Chen, J. S. (2016b). Retrofitting non–cognitive–diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13(3),* 218–230.

Chen, J. S. (2017). A residual–based approach to validate Q–matrix specifications. *Applied Psychological Measurement, 41(4),* 277–293.

Chen, J. S., & de la Torre, J. (2013). A general cognitive diagnosis model for expert–defined polytomous attributes. *Applied Psychological Measurement, 37(6)*, 419–437.

Chen, J. S., de la Torre, J., & Fu, X. (2015, December). *Investigating assumption deviation in the cognitive diagnosis model using a probabilistic process.* Paper presented at the Global Chinese Conference on Educational Information and Assessment & the Chinese Association of Psychological Testing Annual Conference, Taizhong, Taiwan.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76(2),* 179–199.

Doornik, J. A. (2003). *Object–oriented matrix programming using Ox (Version 3.1)* [Computer software]. London: Timberlake Consultants Press.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited–information goodness–of–fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69(3),* 225–252.

Heaton, J. B. (1988). *Writing English language tests* (Vol. 1, pp. 1–13). New York: Longman.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension

ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26(1),* 31–73.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32(2),* 227–258.

Köhn, H. F., & Chiu, C. Y. (2017). A procedure for assessing the completeness of the Q–matrices of cognitively diagnostic tests. *Psychometrika, 82,* 112–132.

Larsen–Freeman, D. (2003). *Teaching language: From grammar to grammaring.* Boston, MA: Thomson/Heinle.

Munby, J. (1981). *Communicative syllabus design: A sociolinguistic model for designing the content of purpose–specific language programmes.* Cambridge: Cambridge University Press.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York: Guilford Press.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20(4),* 345–354.

von Davier, M. (2005). *A general diagnostic model applied to language testing data. Research Report.* ETS RR–05–16. ETS Research Report Series.

Zhan, P. D., Jiao, H., & Liao, D. D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology,* doi: 10.1111/bmsp.12114.

Exploring Cognitive Diagnosis Retrofitting and Further Analyses of Language Proficiency Testing: The Case of the Guangzhou English Achievement Examination

Lin Yanting[1], *Chen Huilin*[2], *Chen Jinsong*[3]

([1]Department of Psychology, Sun Yat-Sen University, Guangzhou, 510006)

([2] School of International Education, Shanghai International Studies University, Shanghai, 200083)

([3] Department of Psychology, Sun Yat-Sen University, Guangzhou, 510006)

Abstract Cognitive diagnostic models (CDMs) can provide meaningful diagnostic information about individuals' knowledge state. Recently, retrofitting CDMs to language tests is increasingly popular. However, existing studies on the topics suffered some issues, largely due to incomplete validation procedures, missing item-level fit measures and superficial analyses. For these reasons, this paper intended to accomplish three tasks. First, it intended to revise validation procedure and strategy based on previous research, and then to verify the validity of the proposed procedure. Second, it intended to retrofit an achievement examination with CDM and to conduct in-depth analyses based on revised validation procedure and strategy. Third, it intended to investigate the language characteristic of English learning among middle school students.

The test materials of this study come from the 2015 Guangzhou Middle-School English Achievement Examination. They include sentence completion and reading comprehension, with about 40 items in total. Data of 2718 students from this examination were analyzed. This research compared two Q-matrixes constructed on the basis of the examination syllabus and expert panel separately, and found that former Q-matrix was less appropriate for cognitive diagnosis. With the revised validation procedures and item-level fit measures, we found that the ability attribute definitions based on the examination syllabus were excessively broad but should be more specific. In comparisons, the attribute set and Q-matrix based on expert panel can be appropriately retrofitted and validated with the procedures and fit measures. Meanwhile, this study further analyzed the retrofitting test and found that: (a) Proficiency classifications based on attributes distribution and total score were different in determining whether a student was passing or not. Whether this was a special case or not can be a topic of further study; (b) The attribute mastery probability showed that student mastery was good in general. The mastery probability of attribute AR3 was the lowest and the hierarchy of attribute AR3 indicated that students need to pay more attention to learning it; (c) There was no significant gender difference on mastering attribute AR4. But there were significant gender differences on the other probability (ts (1, 2716) > -2.51, ps <.012) and girls' level of mastery was significantly better than boys'. Therefore, boys should strengthen their English study; (d) The attribute distribution of attribute patterns of Language Knowledge and Application showed that the attribute profile "11111111" was the largest proportion (29%). The attributes profile "1111" of Reading Comprehension accounts for 23%. It reflected that the relationships among language attributes are interrelated, and provided another evidence for fitness of the G-DINA model in diagnosing the language test or language skills; (e) To test the external validity of our results, the students' listening and writing performance were used as external criteria for evaluation. It showed that the correlations with most attribute probabilities were statistically and substantively significant, suggesting good external validity. In general, this study can lay a foundation of further developing language proficiency testing for cognitive diagnosis purpose.

Key words cognitive diagnosis model, English achievement examination, Q-matrix, G-DINA, fit indices