

中央高校基本科研业务费专项资金资助项目
Fundamental Research Funds for the Central Universities

基于评价使用论证的 英语口语测试效度验证

——以一项初中二年级口语测试体系为例

Building and Validating an
Assessment Use Argument for an English Speaking Test

高 淼 著



中国财经出版传媒集团



经济科学出版社
Economic Science Press

中央高校基本科研业务费专项资金资助项目
Fundamental Research Funds for the Central Universities

基于评价使用论证的 英语口语测试效度验证

——以一项初中二年级口语测试体系为例

Building and Validating an

Assessment Use Argument for an English Speaking Test

高 淼 著

中国财经出版传媒集团



经济科学出版社
Economic Science Press

图书在版编目 (CIP) 数据

基于评价使用论证的英语口语测试效度验证：以一项初中二年级口语测试体系为例/高森著. —北京：经济科学出版社，2016. 6

ISBN 978 - 7 - 5141 - 7087 - 0

I. ①基… II. ①高… III. ①英语 - 口语 - 考试方法 - 研究 - 初中 IV. ①G633.412

中国版本图书馆 CIP 数据核字 (2016) 第 159810 号

责任编辑：王 娟

责任校对：靳玉环

责任印制：李 鹏

基于评价使用论证的英语口语测试效度验证

——以一项初中二年级口语测试体系为例

高 森 著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：010 - 88191217 发行部电话：010 - 88191522

网址：www. esp. com. cn

电子邮件：esp@ esp. com. cn

天猫网店：经济科学出版社旗舰店

网址：http: //jjkxcbs. tmall. com

北京密兴印刷有限公司印装

710 × 1000 16 开 17.75 印张 300000 字

2016 年 10 月第 1 版 2016 年 10 月第 1 次印刷

ISBN 978 - 7 - 5141 - 7087 - 0 定价：48.00 元

(图书出现印装问题，本社负责调换。电话：010 - 88191502)

(版权所有 侵权必究 举报电话：010 - 88191586)

电子邮箱：dbts@ esp. com. cn)

序

检验一个测试项目的质量是确保测试结果有效和有用的必要措施。当前语言测试界较为通用的检验方法是借用巴赫曼和帕默尔 (Bachman & Palmer, 2010) 的“评价使用论证 (AUA)”理论框架, 针对特定考试的测试后效、决策制定、分数解释、测试记录以及考生表现等各个环节收集证据, 验证考试是否考查了预期的语言能力, 并且产生了预期的考试后效。

基于《英语课程标准》的“八年级英语学习成绩检测、分析反馈与指导”项目旨在通过测评评价学生英语语言运用能力发展情况, 为区域基础教育领导教育决策、为一线教师提升教学质量提供有价值的反馈信息。学生和教师问卷调查数据表明, 学生英语口语能力发展被认为是最重要的语言技能也是进步最困难的技能。因此, 通过口试数据, 经过分析获取学生口语表达能力的信息, 并据此提有意义的反馈信息是十分重要的。

高森的这部著作全面细致地报告了研究者构建口语测试 AUA 框架, 收集多种量化和质性相关证据加以分析, 对涉及质量属性的主张和理由进行证明, 从而实现对口语测试的 AUA 验证的系统研究。该研究是国内近年来在语言测试领域所做的较为扎实的研究之一。高森对研究所采用的 AUA 效度验证理论有深刻理解, 并且能够较好地运用于所做研究。专著的出版对于目前语言测试的理论发展与实践是有益的贡献, 同时也为语言测试研究者在研究方法方面做出示范。



2016 年 10 月于北京

前 言

本书旨在基于“评价使用论证 (Assessment Use Argument, AUA)”理论框架 (Bachman & Palmer, 2010), 对“义务教育阶段英语学习质量监测系统”中的初中二年级口语测试进行效度验证。评价使用论证是教育评价领域指导测试开发和效度验证的最新理论范式, 是对基于论证的效度验证 (Kane, 1992, 2002, 2004; Kane, Crooks & Cohen, 1999) 的进一步发展。在这一视角下, 效度验证是一个论证过程, 此论证过程依据“实用推理模型 (Toulmin, 2003)”, 实现从事实到主张的步步推理。考生的“测试表现”是论证所基于的第一个事实, 从“测试表现”出发可获得“测试记录”这一主张, “测试记录”同时转化为“分数解释”这一主张所依据的事实; “分数解释”进一步转化为下一主张——“决策制定”所依据的事实; 基于“决策制定”这一新的事实, 最后做出有关“测试后果”的主张。评价使用论证强调测试的使用, 将使用测试的后果作为设计和开发测试的首要考虑因素, 因此效度验证也是证明测试使用合理性的推理过程, 包括构建评价使用论证框架和收集证据对其进行验证两个阶段。

“义务教育阶段英语学习质量监测系统”的初中二年级口语测试以《义务教育英语课程标准 (2011 年版)》(以下简称《英语课程标准》) 四级有关“说”这一技能的相关要求为依据, 制定了具可操作性的测试任务开发和设计的细目表, 成为设计口语任务和研究口语测试使用的重要指南。本口语测试采用面试型口试形式, 包括日常会话、朗读短文和看图说话三个测试任务。测试目的是了解初中二年级学习者的口语水平, 为教育行政部门制定相关教育决策提供基于测试的证据。基于测试结果所做的决策对学生、教师等测试利益相关者都会产生不同的影响, 因此保证和探究本口语测试的质量至关重要。

本研究首先构建包括测试记录、分数解释、决策制定和测试后果四个主张的口语测试的评价使用论证, 涵盖了“一致性”、“有意义”、“公平



性”、“概括性”、“相关性”、“充分性”、“价值观敏感度”、“均衡性”和“受益度”九个测试质量属性，然后收集多种量化和质性相关证据加以分析，对关于上述质量属性的主张和理由进行证明，从而实现对口语测试的评价使用论证的验证。量化研究主要包括对所制定的分项评分标准的验证，使用分项评分标准重新评分及基于评分结果的一系列研究，对高低水平两组考生在口语产出表现的复杂度、准确度、流利度以及词汇使用的丰富度和复杂度这些语篇特征的测量指标的比较分析。质性研究主要包括对不同水平考生的口试过程进行的会话分析研究和采用即时追溯性口述方法对考生做题过程的探究。

研究参与者主要包括四个群体。第一个群体是从2011年参加英语口语测试的样本中（约900名考生）随机抽取的140位初中二年级学生，分别来自5个省份，他们的口语测试录音是本研究的主要研究对象；第二个群体是参与评分标准制定及完成重新评分任务的4位研究参与人员；第三个群体是使用新的评分标准对24个录音样本进行试评分的三位测试方向研究生；第四个群体是5位参加即时追溯性口述和访谈研究的初三学生。

本研究采用混合研究设计，根据具体的研究问题，综合使用了多种研究方法，主要包括基于项目反应理论的多面 Rasch 研究、基于概化理论的概化研究、即时追溯性口述研究、访谈研究和会话分析研究等。根据数据类型和研究问题，综合使用了多种数据分析工具，如 SPSS (18.0)、Facets (3.58) 和 NVivo (9.0)；同时使用了多种用于录音转写和分析语篇特征的软件，如 Golden Wave, f4, D-Tools 和 P_Lex (2.0)，对口述及访谈的转写文本和口语测试录音转写文本等进行分析。

研究结果表明：（1）所制定的分项评分标准及对其所做的验证为“测试记录”主张及“分数解释”主张从一定程度上提供了直接相关的支持性证据；使用分项评分标准进行重新评分，评分结果与使用整体评分标准所评得分显著相关。（2）对测试分数的概化研究结果表明来自考生水平侧面的方差分量所占比例最大，这一方差分量属于测量目标引起的有效变异，证明测试得分是欲测构念——考生口语能力的体现，为“分数解释”这一主张提供了证据。此外，决策研究表明采用目前的三个口语测试任务，可以实现较好的可靠性系数。（3）多面 Rasch 分析表明测试任务可以较好地地区分不同水平的考生，且评分员评分时有较好的内部一致性，为“测试记录”的一致性提供了支持证据。（4）对考生口试语料的语篇特征的量化研究分析结果表明，高低水平两组考生在准确度、流利度、复杂度



以及词汇的丰富度和难度方面存在显著差异。(5) 会话分析结果表明,不同水平的考生在话轮转换、会话序列发展、修正以及交际的得体度方面存在明显差异,即时追溯性口述的研究结果表明,考生在作答过程中所运用到的能力和关注的方面与欲测构念及有关第二语言口语产出的相关心理语言学理论模型相契合,这些都为基于分数所做的对考生能力推断的有意义度和概括性提供了证据。以上实证研究结果及项目文本材料(即有关口语测试设计的纲领和指导性文件、口试实施手册、考官培训手册等)为证明本口语测试的评价使用论证所包含的主张和相应理由提供了一定的证据。总体看来,支持测试记录及分数解释主张的证据多于支持决策制定和测试后果这两个主张的证据。这是由于收集支持决策制定和测试后果主张的证据的责任更多由决策制定者和测试使用者承担,而笔者参与测试设计,相对来说,对证明测试评分的一致性(测试的信度)和分数解释的有意义、公平性、概括性以及相关性负有责任。从这个角度来看,本口语测试的评价使用论证在较大程度上得到了证明。

本研究主要有以下启示:第一,义务教育阶段英语学习质量监测的初中二年级口语测试的测试目的及测试性质使评价使用论证成为对其进行效度验证的最优途径,本书是在语言测试领域应用这一最新效度研究范式的一个尝试。第二,分为四个维度的分项评判标准可被英语教师直接应用于课堂口语评价,且可基于标准给予学生详细反馈。第三,本研究所使用的包括多种研究方法和统计手段的混合研究设计,相对以往效度验证研究,在方法论上有所突破和深化。第四,对一项测试的评价使用论证的验证是一个持续不断的过程,除了测试设计者应为测试记录及分数解释等主张收集证据进行证明外,测试使用者和决策制定者也应该关注测试使用的影响、决策的均衡性及价值。这些都是未来研究需关注的方面。

目 录

第1章 引言	1
1.1 研究背景	1
1.1.1 基础教育质量监测项目	2
1.1.2 小样本英语口语测试	3
1.2 选题缘起	4
1.3 评价使用论证概述	6
1.4 研究目的与意义	7
1.5 主要术语	9
1.6 本书结构	11
第2章 文献综述	12
2.1 构念及表现性评价中的构念界定	12
2.1.1 构念的定义及界定构念的视角	12
2.1.2 表现性评价的构念界定	14
2.2 效度和效度验证	16
2.2.1 效度概念的变化	16
2.2.2 效度验证的发展	21
2.3 评价使用论证 (AUA)	27
2.3.1 AUA 的基本结构	27
2.3.2 评价使用论证与基于论证的效验的比较及其特点	32
2.4 第二语言口语能力	35
2.4.1 语言能力	36
2.4.2 对第二语言口语能力的认识	40
2.4.3 口语的本质	44



2.5 表现性评价与第二语言口语测试	48
2.5.1 表现性评价	48
2.5.2 口语测试模型	50
2.5.3 相关的第二语言口语测试研究视角	55
2.6 相关研究	62
2.6.1 评价使用论证研究	63
2.6.2 混合研究方法	64
2.6.3 质性研究与口语测试效度研究	64
2.6.4 对口语测试信度的量化研究	68
2.7 本章总结	71
 第3章 英语口语测试的评价使用论证框架	 74
3.1 测试记录	75
3.1.1 主张	75
3.1.2 理由	75
3.2 分数解释	76
3.2.1 主张	76
3.2.2 理由	77
3.3 决策制定	79
3.3.1 主张	80
3.3.2 理由	80
3.4 测试后果	81
3.4.1 主张	81
3.4.2 理由	81
3.5 本章总结	82
 第4章 研究设计	 84
4.1 研究目的和研究问题	84
4.2 方法论及研究设计	85
4.3 参与人员	86
4.3.1 口试考生群体及抽样	86
4.3.2 评分标准制定人员	87
4.3.3 评分员	88



4.3.4 口述报告研究被试	88
4.4 研究工具	89
4.4.1 口试试题	89
4.4.2 分项评分标准	91
4.4.3 即时追溯性口述	92
4.4.4 访谈	92
4.5 研究步骤	93
4.5.1 分项评分标准制定及验证	93
4.5.2 评分	95
4.5.3 口语语料语篇特征分析	96
4.5.4 会话分析	99
4.5.5 做题过程研究	99
4.6 数据分析方法	101
4.6.1 量化数据分析方法	102
4.6.2 质性数据分析方法	112
4.7 方法论意义	115
第5章 研究结果	116
5.1 有关“测试记录”主张的研究结果	116
5.1.1 分项评分标准的验证及修订	116
5.1.2 评分信度研究	130
5.2 有关“分数解释”主张的研究结果	148
5.2.1 口语产出的语篇特征分析	148
5.2.2 会话分析研究	154
5.2.3 考生做题过程研究	164
5.3 本章总结	178
第6章 口语测试评价使用论证 (AUA) 的证据 汇总及讨论	181
6.1 英语口语测试的 AUA 主张	183
6.2 理由、反驳及证据	183
6.2.1 测试记录	184
6.2.2 分数解释	186



6.2.3 决策制定	191
6.2.4 测试后果	191
6.3 本章总结	192
第7章 结语	193
7.1 研究结果	193
7.1.1 研究问题一：支持“测试记录”主张的实证 证据有哪些	194
7.1.2 研究问题二：支持“分数解释”主张的实证 证据有哪些	194
7.2 研究价值及启示	196
7.2.1 效度验证的理论与实践方面	196
7.2.2 对口语测试实践的指导方面	197
7.2.3 研究方法方面	197
7.2.4 口语教学方面	198
7.3 研究局限及前景	198
附录1：整体评分标准	200
附录2：口试测试试题及对三组图片的详细描述	201
附录3：做题过程后的追踪访谈提纲	203
附录4：分项评分标准	204
附录5：会话分析转写符号体系	206
附录6：口试录音语料转写举例	208
附录7：被试口述报告及访谈录音转写文本举例	210
附录8：NVivo 编码形成过程	217
附录9：各维度上各侧面的具体方差分量估计值汇总	221
附录10：每位被试的编码节点情况	224
参考文献	227
后记	262

表 目 录

表 1-1	2011 年初中二年级口语测试试卷结构	3
表 2-1	界定构念的七种基本视角	13
表 2-2	早期效度概念	17
表 2-3	分层效度框架	18
表 2-4	对完形填空测试使用的效度证据的分析框架	20
表 4-1	抽样情况	87
表 4-2	口述研究被试情况	89
表 4-3	《英语课程标准》与口语测试指标对照	90
表 4-4	研究数据收集汇总	102
表 4-5	概化研究方差分量计算公式列表	104
表 4-6	语篇特征分析指标	110
表 4-7	研究方法汇总	114
表 5-1	评分员评分情况(多面 Rasch 分析)	117
表 5-2	考生能力(多面 Rasch 分析)	118
表 5-3	评分量表的四个维度统计结果	119
表 5-4	任务完成度维度的分数段统计	121
表 5-5	语音语调维度的分数段统计	122
表 5-6	语言准确性维度的分数段统计	124
表 5-7	交际有效性评分量表分数段统计	125
表 5-8	修订后的分项评分标准对各任务的评分维度	127
表 5-9	新旧评分标准相关研究描述性统计量	128
表 5-10	新旧评分标准相关统计量	128
表 5-11	各维度方差分量估计值及所占百分比	130
表 5-12	四个维度随机双面交叉设计 $p \times I \times R$ 决策研究结果	133



表 5-13	考生的能力 (多面 Rasch 分析)	135
表 5-14	评分员的评分情况	136
表 5-15	偏差分析: 评分员与试题的交互作用	138
表 5-16	任务难度指标	139
表 5-17	评分量表四个维度的统计量	140
表 5-18	评分量表分数段统计	141
表 5-19	评分量表 (语音语调维度) 分数段统计	142
表 5-20	评分量表 (准确性维度) 分数段统计	144
表 5-21	评分量表 (交际有效性) 分数段统计	145
表 5-22	评分量表 (任务完成度) 分数段统计	146
表 5-23	高、低两组考生得分的描述统计	149
表 5-24	高、低两组考生得分的均值检验统计结果	149
表 5-25	口语产出总词数描述统计结果	149
表 5-26	高、低两组考生口语产出总词数的均值 检验统计结果	150
表 5-27	任务三的口语产出语篇分析的描述统计结果	150
表 5-28	高分组和低分组考生在不同测量指标上的 均值比较结果	151
表 5-29	词汇统计指标描述数据分析结果	152
表 5-30	两组被试词汇指标均值检验统计结果	153
表 5-31	词汇复杂度进一步研究的描述统计量	153
表 5-32	词汇复杂度进一步研究的均值检验统计量	153
表 5-33	被试对图片信息的关注	170
表 5-34	被试在口试作答中遇到的困难	175
表 5-35	“努力揣测图片包含的信息” 的编码节点	177
表 6-1	主张、理由与证据	182

图 目 录

图 1-1	评价使用论证框架	6
图 2-1	解释性论证推理链	24
图 2-2	图尔明的实用推理模型	24
图 2-3	新托福的解释性论证推理链	25
图 2-4	评价使用论证 (AUA) 推理过程	28
图 2-5	评价使用论证框架	33
图 2-6	语言能力的组成部分	39
图 2-7	表现性评价框架	49
图 2-8	口语测试框架	51
图 2-9	口语测试表现框架	53
图 2-10	扩展后的口语测试表现框架	54
图 2-11	口语测试模型	55
图 2-12	EBB 评分标准图示 (Turner & Upsher, 2002)	60
图 2-13	树形评分标准 (Performance Decision Trees)	61
图 3-1	口语测试的评价使用论证框架及验证证据来源	83
图 4-1	分项评分标准制定的过程	93
图 5-1	评分量表分数段的概率曲线图 (任务完成度维度)	121
图 5-2	评分量表分数段的概率曲线图 (语音语调维度)	123
图 5-3	评分量表分数段的概率曲线图 (语言准确性维度)	124
图 5-4	评分量表分数段的概率曲线图 (交际有效性维度)	126
图 5-5	新旧评分标准相关研究散点图	129
图 5-6	多层面 Rasch 模型分析结果	134
图 5-7	分数段的概率曲线图	141



图 5-8	评分量表分数段的概率曲线图 (语音语调)	143
图 5-9	评分量表分数段的概率曲线图 (准确性)	144
图 5-10	评分量表分数段的概率曲线图 (交际有效性)	146
图 5-11	评分量表分数段的概率曲线图 (任务完成度)	147
图 5-12	三级编码核心类属之间的关系模型图	166
图 5-13	被试 3 对“我的好朋友 Linda”组图的描述	171
图 5-14	被试 4 对“Linda 是我的好朋友”组图的描述	171
图 5-15	被试 2 对“Tom 的一天”组图的描述	172
图 5-16	被试 5 对“Women's Day”组图的描述	173

学习一门外语的主要目标是会“说”这门外语，但这却是众多语言学习者最难实现的。在语言测试领域，相比其他技能（听、读、写）的测试，口语测试的设计和实施都比较困难，甚至“口语能力”这一构念究竟所指为何还存在诸多争议。因此，在口语测试的设计中，清晰地界定构念，进而有效评估考生的口语能力，以此促进学生口语学习的发展，是一项非常重要的研究任务，而效度验证是保证测试质量之必须。本研究基于“评价使用论证（Assessment Use Argument, AUA）（Bachman & Palmer, 2010）”这一效验理论框架，对“义务教育阶段英语学习质量监测”系统中的初中二年级口语测试进行效度验证。本部分主要从研究背景、选题缘起、理论概述、研究目的与意义以及术语解释和内容结构这六个方面进行介绍。

1.1 研究背景

随着信息技术的发展和数据模式的爆炸式增长，当今世界已经进入网络化的“大数据时代”，美国等发达国家已将对“大数据”的研究上升至战略高度（李和程，2012；郭，2012）。在教育领域，构建和发展“教育大数据”对于制定教育政策、优化教育制度、促进教师教学、提升学习质量意义重大。长期的、大规模的、科学的教育质量测评是收集“教育大数据”的重要工具，如国际学生评价项目（PISA，国际经合组织 OECD 主持），美国国家教育进步评价项目（NAEP，美国国家评价管理委员会 NAGB 主持）以及国际数学与科学趋势研究项目（TIMSS，国际教育成就评价协会 IEA 主持）。我国义务教育阶段学习质量监测体系即创建和发展



于这一时代背景下,该体系通过在全国范围内对四年级、初中二年级学生进行全方位的测试,获取有关中国基础教育质量以及影响因素的大量数据,从而为决策制定提供依据和建议。本部分介绍本监测项目的实施背景以及测试工具的内容和特征。

1.1.1 基础教育质量监测项目

我国基础教育质量监测项目(以下简称“质量监测”)始于2007年,通过对处于义务教育阶段的学生的学习质量、身心健康状况、艺术素养、实践能力和创新意识的测量,以及对影响学生发展的相关因素的监测,为国家教育决策的制定提供信息、依据和建议。

监测对象为四年级和初中二年级学生。选择这两个年级作为监测对象,主要基于以下两点考虑:第一,参照了国际上比较有影响力的教育测评项目(PISA, NAEP, TIMSS),这些项目的测试对象多为四年级和初中二年级的学生,PISA更明确其测试对象为15岁学生。在我国,四年级和初中二年级均处于义务教育阶段,对该阶段的教育教学情况以及学生的学习质量进行监测是国家教育发展之所需。此外,我国学生完成初中二年级学习时年龄通常在15周岁左右,所以监测结果与国际其他监测项目具有较大的可比性。第二,选择初中二年级学生作为监测对象,是考虑该年级处于初中学习的中间阶段,相比初中一年级学生来说,学生在心理及情感态度方面已经基本适应中学学习;学习内容也在难度和广度方面有了较大提高。而初三学生由于面临中考,备考压力难免会对教学和学习产生一系列影响(Alderson & Wall, 1993; Bailey, 1996; Cheng & Watanabe, 2004; Hawkey, 2006; Shohamy, Donitsa - Schmidt & Ferman, 1996),因此监测结果可能难以如实反映正常教育环境下的教学和学习以及其他相关因素的情况,所以,监测对象确定为四年级和初中二年级这两个学生群体。

2011年9月,中华人民共和国教育部基础教育质量监测中心开展了对全国11个省(自治区、直辖市)的监测,对被抽样的四年级和初中二年级学生进行了英语学习质量(纸笔测试)、体育学力与体质健康状况、学校与家庭相关因素等方面的监测。2011年10月份,对北京、吉林、青海、广西和福建5个地区的四年级、初中二年级考生进行了口语测试,各个年级的抽样人数均为900人。相对于大规模的笔试测试,将这一测试称之为