文章编号:1004 - 5139(2022)03 - 0082 - 11 中图分类号:H059 文献标识码:A

数据驱动下的口译分项评估权重研究:母语为汉语的评分员视角

尚小奇[1],李德超[2]
(1. 深圳大学 外国语学院,深圳 518060; 2. 香港理工大学 中文及双语学系,香港)

摘 要:口译分项评估是目前应用较为广泛的口译质量评估方式。然而,关于评估参数权重的设定,口译研究者之间并未达成共识。现有研究大多通过理论推演或问卷调查等方法来探讨权重的分配,而基于评分员真实评分、以数据为驱动的实证研究仍然比较匮乏。在探讨口译权重设定时,现有研究也鲜将方向性考虑在内。鉴于此,本研究以英汉双向口译为研究对象,通过分析八位评分员(母语:汉语;外语:英语)对 50 个口译学员(母语:汉语;外语:英语)的口译录音的评分数据,以探讨如何在各个参数间分配权重。数据分析结果显示:(1)无论何种语言方向,信息始终是最为重要的评估参数(β_1 = .351(汉英); β_1 = .593(英汉));(2)无论何种语言方向,表述的权重均位列第二(β_3 = .345(汉英); β_3 = .381(英汉));(3)语言在汉英口译评估时的权重排序第三(β_2 = .325),而对于英汉口译而言,其权重无法通过统计模型估算出来。本研究的发现可以为英汉双向口译评估标准的设定和口译培训提供重要实证数据。
关键词:英汉口译;分项评估;权重;方向性

A Data-driven Approach to Exploring Weighting Schemes for Assessing Bi-directional Interpreting Performance: Evidence from Native Chinese-speaking Raters

SHANG Xiaoqi[1], *LI Dechao*[2]
(1. School of Foreign Languages, Shenzhen University, Shenzhen 518060, China; 2. Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China)

Abstract: Analytic rating scales are widely used for assessing interpreting. However, weighting schemes for assessing interpreting reported in previous studies have been largely conceptual and generally pre-determined. Research that investigates weighting based on empirical interpreting assessment data remains scant. And few studies to date have attempted to differentiate between language directions when it comes to weighting. To fill this gap, this study adopts a data-driven approach to exploring weighting schemes for assessing Chinese to English bi-directional interpreting performance. A total of eight raters were invited to evaluate 50 Chinese to English (C-E) interpretations and 50 English to Chinese (E-C) interpretations by trainee interpreters, using an analytic rating scale and a holistic rating scale. Data analysis suggested that: (1) fidelity was the predominant criterion in predicting the candidate's interpreting performance, regardless of interpreting direction (β_1 = .351 for C-E interpreting; β_1 = .593 for E-C interpreting); (2) delivery came second among the three assessment criteria, regardless of interpreting direction (β_3 = .345 for C-E

interpreting; β_3 = .381 for E-C interpreting), and (3) language contributed up to 32.5 percent (β_2 = .325) of the variance in the candidate's interpreting performance in the C-E direction, whereas its predictive power on interpreting performance failed to be detected in the E-C interpreting direction due to statistical concerns. Implications of the findings for interpreter training and for the development and validation of assessment tools for interpreting performance are discussed at the end.

Key words: Chinese-to-English interpreting; analytic rating; weighting scheme; directionality

1. 引言

评估(assessment)在语言测试领域有着至关重要的作用,有效的评估有助于甄选合格的学员、跟踪和改进语言学习和教学以及做好职场准入的守门人(Bachman 1996; Weigle 2002; Weir 2005)。同样,在口译培训领域,评估也服务于多种目的,如项目准入(program entry)、中期分流(mid-streaming)和学位授予(degree conferral)等(Sawyer 2004),有效的评估可以提高口译培训成功的概率(Roy 1984)。

多年来,口译质量评估的参数一直是口译研究者关注的焦点之一(Bühler 1986; Ng 1992; Schjoldager 1995; Clifford 2001; Pöchhacker 2001; Riccardi 2002; J. Lee 2008; Skaaden 2013; S-B. Lee 2015)。尽管这些研究者设计的参数数量各不相同,但其核心参数主要涉及三个方面,即信息忠实度、目的语质量和表述流畅度。然而,对于如何在这三者之间合理地分配权重,研究者之间仍然存在着分歧(如 Roberts 2000; Riccardi 2002; J. Lee 2008; Choi 2013; Skaaden 2013; Wu 2013; S-B. Lee 2015)。现有关于口译评估参数权重的研究大多为理论、概念推演或基于对译员的大规模问卷调查,而基于真实评估数据的实证研究仍然非常匮乏。在探讨权重时,目前鲜有研究将"方向性"(directionality)考虑在内。而相关研究表明,方向性会对口译策略和口译产出产生影响(如 Darò *et al.* 1996; Tommola & Heleva 1998; Donovan 2005; Chang & Schallert 2007; Setton & Dawrant 2016)。如 Setton & Dawrant (2016:219)认为,当译入外语时,口译员具有"理解优势"(comprehension bonus),但会面临"产出赤字"(production deficit);而相反,当译入母语时,口译员则具有"产出优势"(production bonus),但会面临"理解赤字"(comprehension deficit)。鉴于此,笔者认为,在评估不同译入方向的口译产品时,评估者对不同评估参数的侧重也可能会有所不同,因此有必要开展基于实证评估数据的口译参数权重的研究,而且亦将方向性这个变量考量纳入考察范围。

本研究以数据为驱动,以英汉和汉英口译(交替传译)为研究对象,通过统计分析八位评分员(母语为汉语,外语为英语)对 50 个口译学员(母语为汉语,外语为英语)口译录音(英汉和汉英各 50 个)的评估数据,试图探讨如下两个研究问题:

(1) 在评估英汉和汉英口译时,评分员是否会对不同评估参数有不同的侧重?

(2) 如果有不同侧重,其背后的原因可能是什么?

2. 文献综述

2.1 口译评估的参数

"分项评估"(analytic rating)是在语言测试领域被广泛采用的一种评估方式。和"整体评估"(holistic rating)关注于整体表现不同,分项评估要求评估者从各个参数对受试者表现进行评估,可以提供更多具有信度和效度的数据(Green & Hawkey 2010; Weigle 2002)。

多年来,口译研究者通过理论探讨(Schjoldager 1995; Roberts 2000; Riccardi 2002; Skaaden 2013)、问卷调查(Bühler 1986; Ng 1992; Kurz 2002a, 2002b)或实验研究(Clifford 2005; J. Lee

2008；Choi 2013；Wu 2013；S-B. Lee 2015）等多种方法来探讨口译评估的参数。这些研究中口译参数的数量包含三个宏观参数（Clifford 2005；J. Lee 2008；Choi 2013）到 17 个微观参数（Riccardi 2002）不等。如 Schjoldager（1995）制定了同声传译的质量评估“清单”（checklist），主要涵盖信息忠实度、语言质量、连贯和表述流畅度四个宏观参数，每个参数均包含 5 个以上的子参数。Riccardi（2002）的研究则使用多达 17 个参数，如词汇偏离、语言偏离等。尽管使用的术语有所区别，但口译研究者对三个宏观参数达成了共识，即信息忠实度、目的语质量和表述流畅度。

2.2 口译评估参数的权重

口译研究者对于如何分配参数的权重存在着分歧。有学者主张采取等量齐观的做法（Cheung 2007），也有学者认为不同参数应该具有不同的权重（如 Roberts 2000；J. Lee 2008；Choi 2013；Skaaden 2013；Wu 2013；S-B. Lee 2015）。

有的学者通过文献研究和理论推演来探讨口译参数的权重。J. Lee（2008）通过整合相关文献，在评估英－韩交替传译的研究中将信息忠实度、目的语质量和表述流畅度的权重分别设定为 40%，40% 和 20%。而同样在评估英－韩交替传译时，Choi（2013）则将三者的权重设定为 50%，30% 和 20%。

也有学者基于实证评分数据来探讨不同参数的权重。S-B. Lee（2015）通过统计分析两位评分员对 33 位口译学员的英－韩交替传译评分数据发现，信息忠实度的权重应为 50%，而目的语质量和表述流畅度的权重则均为 25%。Wu（2013）则通过对 30 位评分员关于 5 位口译学员的英汉同声传译录音评估数据的分析，发现在 5 个评分参数中，信息忠实度和表述流畅度占据最为核心的权重。

问卷调查也被用于探讨口译参数的权重（Bühler 1986；Kurz 2001，2002；Pöchhacker 2012）。Bühler（1986）围绕 15 个“语言内”（linguistic）和“语言外”（extra-linguistic）参数，对 47 名国际会议口译协会（AIIC）会员开展了问卷调查。研究发现，“意义的一致性”（sense consistency）被认为是最为重要的评估参数，术语的准确性和流畅度次之，而口音则被看作最为次要的参数。同样，Pöchhacker（2012）基于对 704 名 AIIC 会员的问卷调查也发现，在参数的重要性方面，位列前三的参数依次为意义的一致性、逻辑的衔接（logical cohesion）和流畅度。

从如上文献可以看出，口译研究者更多是通过理论推演或问卷调查来探讨不同参数的“名义权重”（nominal weight），倾向认为信息忠实度是评估口译质量时最为重要的参数，而对于其他两个核心参数（目的语质量和流畅度）的看法却存在着分歧。由于有很少证据表明“名义权重”等同于“实际权重”（effective weight）（Wang & Stanley 1970），因此，以数据为驱动的、基于评分员的现实评分数据对不同参数权重进行探讨的研究就变得非常必要。除 Han（2015）探讨了英汉双向口译中权重的差异外，现有的基于评分数据的实证研究大多关注单一的语言方向（如：Wu 2013；S-B. Lee 2015），“方向性”对口译评估参数权重的潜在影响并未得到应有的关注。

2.3 口译的方向性

“方向性”在本研究中指的是口译员在口译时涉及的包括第一语言（L1）和第二语言（L2）的语言方向。L1 是口译员处于“支配地位”（dominant）的语言，通常是其“母语”或“本族语”（Pokorn 2007），L2 是口译员处于非支配地位的语言，通常通过其后天习得（Crystal 2008：321）。

研究表明，方向性会影响口译策略的使用（如 Bartlomiejczyk 2006；Chang & Schallert 2007），从而影响口译产出的质量（如 Tommola & Heleva 1998；Y-H. Lee 2003；Donovan 2005；Chang & Schallert 2007；Setton & Dawrant 2016）。如 Bartlomiejczyk（2006）对比了汉英双向口译时译员采取的不同策略。研究发现，相对于译入 L1 而言，在译入 L2 时，译员采用更多的概括和整合策略。

Y-H. Lee（2003）通过对比职业译员韩－英双向口译的产出发现，和译入 L1 相比，译员在译入 L2 时会产生较多语言错误但较少信息方面的错误。相反，Tommola & Heleva（1998）关于英语－芬兰语双向口译质量的研究则发现，学生译员在不同语言方向的口译产出没有显著性差异。

鉴于口译方向性对口译质量的重要性以及口译参数权重研究存在的不足，本研究基于评分员的实证评分数据，试图探讨在评估汉英双向交替传译时，评分员是否对不同口译参数的权重有所不同以及其背后的成因。

3. 实验设计

3.1 受试者

本研究选取来自于国内四所高校的 50 名翻译专业硕士（MTI）项目的一年级学生为受试对象，其中女生 43 名，男生 7 名，年龄介于 22～24 岁之间。所有受试者的 L1 均为汉语，L2 均为英语。实验在第一学年年末开展，所有受试者均已接受了系统、完整的口译培训，掌握了口译的基本技能。

3.2 实验材料

本实验包含汉英和英汉双向交替传译测试。汉语演讲主题为“中国烟草管理”，选自国内某知名院校口译培训中期分流测试真题。测试材料为视频，讲者母语为汉语，时长约 4 分钟，语速约每分钟 175 个字。

英语演讲主题为“中国环境保护”，选自欧盟口译训练视频库 Speech Repository。测试材料也为视频，讲者母语为英语，时长约 5 分钟，语速约每分钟 170 个音节。

3.3 实验流程

本实验分别于 2018—2019 学年末在受试者所在大学的语音实验室开展。实验顺序为中译英和英译中测试，每个测试均包括三个部分。测试开始前，受试者会被提供一份与测试相关的主题知识和词汇表，并有 15 分钟的译前准备时间。测试期间，受试者在听到每一部分的视频演讲后，将其口译为目的语。口译期间受试者允许做笔记。测试过程全程录音，测试后研究者收集所有口译录音，并将其编码。

3.4 数据收集

3.4.1 评分标准

本研究的口译评分标准借鉴了 Setton & Dawrant（2016）和 Schjoldager（1995）的质量评估量表和清单，并做了适当修正。评分参数包括信息忠实度（信息）、目的语质量（语言）和表述流畅度（表述）三个维度。在该研究中，信息指的是完整、忠实并准确地传达源语的信息。语言指的是目的语质量，包括语音、语法、术语、措辞、风格、语域和表达的地道性等。表述指的是产出的可听性和交际性，最小化的使用填充词、停顿、修正等。

评分员需要根据受试者的表现，按照差（1～2）、较差（3～4）、一般（5～6）、良好（7～8）和优秀（9～10）五个等级对每个参数进行评分；此外，评分员按照同样的等级标准对受试者口译的整体表现进行评分。除给出量化等级分数以外，评分员也要对受试者口译表现的优点和不足做出质性评价，以用于数据的三角验证。

3.4.2 评分员培训

本研究共选取八位评分员进行评分。八位评分员的年龄介于 28～36 岁不等，5 位女性，3 位男性；第一语言均为汉语，第二语言为英语；4 位为中国高校口译教师，4 位为专职译员。八位评分员均具有超过 4 年的口译经验。

八位评分员共同参加了时长约 6 个小时的评分员培训。其中汉英口译和英汉口译各 3 个小

时。培训内容包括熟悉测试任务和评分标准、审阅不同等级的“基准”(benchmark)测试样本以及小组讨论和答疑等。

在正式评分前,所有评分员均对20个口译录音进行了预评分(汉英和英汉方向各10个),预评分持续时长为3小时。在完成预评分后,根据“多层面Rasch模型”(MFRM)关于评分员严厉度和评分自身一致性的数据,笔者对个别评分员(如评分员8)进行了再培训,并根据评分员反馈,对评分做了“定锚”(anchoring)。随后八位评分员开始正式评分。

3.5 数据分析

如前所述,在预评分阶段,笔者使用Facets 3.57.0 (Linacre 2005)对八位评分员的评分数据做了“多层面Rasch模型”分析,以探讨评分员的严厉度(logit值)和自身一致性(infit值)。笔者也通过SPSS 22探讨了评分员间信度(组内相关系数,Intraclass Correlation Efficient, ICC)。

在收集到正式评分的数据以后,笔者使用SPSS 22对受试者各维度的成绩和口译的整体成绩做了皮尔逊相关分析,以探讨它们之间的相关性;并基于相关性,对各参数的成绩和口译的整体成绩做了多元线性回归分析,以探讨各个参数对口译整体成绩的预测值,从而获得其所占的“实际权重”。

4. 结果

4.1 预评分信度

Rasch模型分析显示,对于汉英口译评分而言,评分员1~7的logit值为-0.67~+0.98不等,介于-1~+1的可接受区间;infit值为0.56~1.06,介于0.5~1.5的正常区间(Linacre 2005)。8号评分员的logit值为-1.13,infit值为1.98。这说明除了8号评分员,其他7位评分员的评分严厉度和自身一致性均处于可接受的范围。随后笔者对8号评分员进行了再培训。对于英汉口译预评分而言,八位评分员的logit值介于-0.88~+0.98之间,infit值为0.63~1.36不等。这说明八位评分员评分的严厉度处于可接受范围且自身一致性较高。

SPSS统计数据显示,对于汉英口译预评分而言,总分以及各维度分数的ICC值介于0.75~0.92之间,对于英汉口译预评分而言,总分以及各维度分数的ICC值介于0.80~0.90之间,由于0.7以上的ICC值表明评分员间信度较高(Shohamy 1985),因此可以确定本研究中8位评分员具有较高的评分员间信度。

4.2 正式评分

4.2.1 相关分析

表1是汉英口译评分的描述性统计数据。信息、语言、表述的成绩和口译总成绩的均值分别为6.759、6.463、6.318和6.481,标准差介于0.942到1.170不等。

表1 描述统计量(汉英)(n=50)

	均值	标准差	样本量
口译总成绩	6.481	1.140	50
信息	6.759	0.942	50
语言	6.463	1.106	50
表述	6.318	1.170	50

表2是汉英口译各参数成绩和总成绩之间的相关关系数据。如图所示,信息、语言、表述和口译总成绩之间的相关系数分别为 $r_1 = .940$, $r_2 = .964$, $r_3 = .957$。由于当两者的相关系数大于0.7时,为显著相关(许宏晨2013:64),我们可以判定三个参数的成绩和口译总成绩之间均呈显著相关关系。

表2　口译各参数成绩和口译总成绩相关矩阵(汉英)(n=50)

变量	相关矩阵		
	1	2	3
口译总成绩	0.940	0.964	0.957
1. 信息	—	0.891	0.867
2. 语言		—	0.943
3. 表述			—

$*p < .05$

表3是英汉口译评分的描述性统计数据。信息、语言、表述的成绩和口译总成绩的均值分别为7.368、7.378、7.559和7.348,标准差介于0.829到0.977不等。

表3　描述统计量(英汉)(n=50)

	均值	标准差	样本量
口译总成绩	7.368	0.913	50
信息	7.378	0.977	50
语言	7.559	0.829	50
表述	7.348	0.977	50

表4是英汉口译各参数成绩和总成绩之间的相关关系数据。如图所示,信息、语言、表述和口译总成绩之间的相关系数分别为 $r_1 = .955$, $r_2 = .923$, $r_3 = .919$。同样,我们可以判定三个参数的成绩和口译总成绩之间也呈显著相关关系。

表4　口译各维度成绩和总成绩相关矩阵(英汉)(n=50)

变量	相关矩阵		
	1	2	3
总成绩	0.955	0.923	0.919
1. 信息	—	0.862	0.812
2. 语言		—	0.918
3. 表述			—

$*p < .05$

4.2.2 回归分析

回归分析的一个重要前提是自变量之间不存在“共线性”(collinearity)。当自变量的“容差”

(Tolerance)大于(1 -“校正(adjusted)” R^2)时,则表明它与其他自变量之间不存在多重共线性(许宏晨2013:83)。从如下表5和6关于Tolerance的数据来看,均大于(1 -“校正(adjusted)” R^2),故本研究的自变量之间不存在共线性,因此回归方程比较可靠。

如表5所示,多元线性回归分析结果显示:对于汉英方向而言,三个参数(自变量)作为一个整体可以解释口译总成绩(因变量)变异的97.2%(adjusted R^2 = 0.972)。具体而言,信息、语言和表述分别可以解释变异的35.1%(β_1 = .351)、32.5%(β_2 = .325)和34.5%(β_3 = .345)。

表5 回归分析统计表(汉英)(n=50)

变量		R	R^2	Adjusted R^2	F	Beta	t	Sig.	Tolerance	VIF
因变量	口译总成绩	.987[a]	.974	.972	562.577		-3.376	.002		
自变量	1. 信息					.351	6.528	.000	.199	5.013
	2. 语言					.325	4.032	.000	.089	11.273
	3. 表述					.345	4.704	.000	.107	9.340

$*p<0.05$

表6显示,对于英汉方向而言,三个参数(自变量)作为一个整体可以解释口译总成绩(因变量)变异的97.2%(adjusted R^2 = 0.972)。具体而言,信息和表述分别可以解释口译总成绩变异的59.3.1%(β_1 = .593)和38.1%(β_2 = .381)。语言对口译总成绩的预测性无法估算出来($p>.05$)。

表6 回归分析统计表(英汉)(n=50)

变量		R	R^2	Adjusted R^2	F	Beta	t	Sig.	Tolerance	VIF
因变量	口译总成绩	.987[a]	.974	.972	533.874		.700	.000		
自变量	1. 信息					.593	12.116	.000	.254	3.933
	2. 语言					.062	.865	.392	.117	8.536
	3. 表述					.381	6.097	.000	.156	6.424

$*p<0.05$

5. 讨论

从表5和6的回归分析数据,可以看到如下趋势:(1)无论是汉英还是英汉口译,信息均在三个参数中占据最高的权重,可以最大程度地解释受试者口译整体表现的变异。(2)无论是汉英还是英汉口译,表述所占的权重均在三个参数中排名第二。(3)对于汉英口译而言,语言所占的权重为32.5%;而对于英汉口译而言,语言所占的权重无法通过统计分析估量出来。

5.1 信息

本研究关于信息参数权重的发现进一步印证了前人研究的结果(如 Bühler 1986; Jones 1998; Gile 2009; Liu 2015; S-B. Lee 2015; Setton & Dawrant 2016). 如 Setton & Dawrant (2016:434)认为,

对于任何语言方向而言,在语言质量和表述参数的表现可以接受的情况下,信息应该始终被视作最为核心的参数。

然而,值得注意的是,根据本研究得到的数据,尽管对于两个语言方向而言,信息在三个参数中均占据最高的权重,但其在英汉口译中所占的权重(β_1 = .593)要远远高于在汉英口译中的权重(β_1 = .351)。产生这一差别的原因可能是由于受试者和评分员的母语和外语因素所致。如前所述,Setton & Dawrant (2016)认为,在译入外语时,口译员会面临"产出赤字",但具备"理解优势";相反,在译入母语时,口译员具备"产出优势",但会面临"理解赤字"。因此,在译入母语时,由于口译产出的信息维度更多的取决于口译员对于源语的理解,而因源语为其外语,口译员在信息维度的表现则比译入外语时更能区分其口译能力的差别。对于本研究而言,由于评分员的母语和外语背景与口译员相似,在评估译入母语时,评分员也可能会相应的给予信息维度更高的权重。

5.2 表述

本研究数据显示,无论对于汉英和英汉两个语言方向而言,表述所占的权重在三个参数中均位列第二(汉英:β_3 = .345;英汉:β_3 = .381)。这一研究发现也和前人的研究结果一致(如 Bühler 1986; Pöchhacker 2012; Wu 2013)。如前所述,Bühler (1986)对国际口译协会 (AIIC) 会员的问卷调查显示,表述是继信息("意义的一致性""术语准确度")之后最为重要的评估参数。同样,Pöchhacker (2012)对 AIIC 也开展了问卷调查,研究结果也显示,表述紧随"意义的一致性"和"逻辑的连贯"之后,在所有维度中占据重要的权重。

然而我们也要看到,本研究关于表述权重的数据又和 J. Lee (2008)和 Choi (2013)的研究发现不相吻合。J. Lee (2008) 的研究表明,在评估英 – 韩口译时,表述所占的权重最小(20%),而信息和语言两个参数分别被赋予 40% 的权重。在 Choi (2013)的研究中,表述这一参数同样被给予最小的权重(30%)。造成这一差距的原因可能是由于二者研究方法的不同。J. Lee (2008)和 Choi (2013)的研究主要为理论探讨和概念推演,权重大多为"预先设定"(pre-determined);而本研究的结果则基于八位评分员的真实评分数据。由于语言测试领域大量的研究表明,基于实证评分数据的研究比基于直觉驱动的研究更能产生可信的数据(Fulcher 2003; Knoch 2009),未来需要开展更多采用更大样本的、以数据为驱动的实证研究以进一步验证本研究的发现,从而得出更加有说服力的结论。

本研究结果亦显示,对于英汉口译而言,信息和表述的权重之和超过 90%(β_1 = .593,β_3 = .381),这一研究发现和 Wu (2013)关于英汉口译质量评估的研究结果一致。Wu (2013)的研究发现,信息和表述占据全部参数权重的 86%。这两个研究结果相似的原因可能是由于其研究对象均为英汉口译,进一步表明了评分员在评估口译质量时可能会受到语言方向因素的影响。

5.3 语言

如前所述,本研究结果显示,对于汉英口译而言,语言参数可以解释口译总成绩变异的 32.5%(β_2 = .325),而对于英汉口译而言,语言参数对口译总成绩的预测性统计模型无法估量出来。但需要指出的是,尽管在评估汉英口译时,语言参数的权重位列第三,但它和信息维度的权重(β_1 = .351)也相差不多。这说明在译入外语时,口译员的外语水平对其口译整体表现的重要影响,从而相应的也会对评分员产生影响。

对于英汉口译而言,语言维度的预测性通过统计模型无法估量的原因也可能在于方向性。如前所述,当译入其母语时,口译员会经历"理解赤字",但具有"产出优势",因此对于本研究而言,由于汉语是其母语,受试者在语言(汉语)维度的产出表现较难区分其口译的整体表现,评分员可能更多的是从信息和表述维度来判断口译的质量,因而给予语言维度的权重可能无法准确估量。

5.4 对口译质量评估和培训的启示

综上,本研究对于口译质量评估和口译培训有着如下参考价值:

本研究实证数据显示,在评估英汉双向口译时,母语为汉语且为口译员出身的评分员对不同参数有着不同的侧重。方向性对评估产生了一定的影响。这一研究数据为进一步完善口译质量评估的框架提供了重要的实证数据。

首先,对于口译质量评估而言,无论对于何种语言方向,信息维度应该始终被评分员视为最为重要的评估维度。口译的终极目标即准确传达讲者想要传达的信息(Gile 2009; Setton & Dawrant 2016),任何随意增添、删减或扭曲源语信息的做法都不应被允许。

其次,表述在评估英汉双向口译时的重要性应该得到进一步的凸显。由于流畅的表述可以产生良好的"包装效果"(packaging effect)(Gile 2009),以及在某种程度上"讲话的方式比内容更为重要"(Brennan & Williams 1995),"听众导向"(listener orientation)和"目标语文本的可理解性"(target text comprehensibility)(Pöchhacker 2001)的理念应该贯穿口译培训的始终。

再次,在评估汉英口译时,外语产出至关重要。如前所述,对于汉英口译而言,尽管语言的权重在三个参数中位列最后,但仍然和其他两个参数相当。鉴于此,口译培训也应该设计专门的语言提高课程模块或中外教联合教学以监控和提高学员的外语语言产出质量。

6. 结语

本研究是以数据为驱动的针对英汉双向口译质量评估的实证研究。研究结果显示,无论对于哪个语言方向,信息始终被视作最为重要的维度,占据最高的权重;表述位列第二;语言维度对于汉英口译而言其权重为34.5%,在三个维度中排序第三。而对于英汉方向而言,其权重通过统计模型无法估量出来。本研究的发现对于英汉双向口译质量评估框架的设定和口译培训有着重要的参考价值。

本研究也有不足之处:1)由于受试者(n=50)和评分员(n=8)样本较小,本研究的结论仅为试探性,未来研究可以通过开展更大规模的评分,以进一步提高研究结果的普适性。此外,本研究邀请的评分员母语均为汉语,未来研究也可以邀请母语为英语且为译员出身的评分员参与汉译口译评估,以探讨评分员的语言背景差异对其权重分配的潜在影响。2)该研究为量化研究,由于单纯的量化研究往往"过于简单"(over-simplistic)且"去语境化"(decontextualized)(Brannen 2005:7),未来的研究可以采用混合研究方法,通过有声思维或访谈获取更多的质性数据,以对量化数据进行三角验证,进一步提高研究结果的信度。

参考文献:

[1] Bachman, L. F. & A. Palmer. *Language Testing in Practice* [M]. Oxford: Oxford University Press, 1996.

[2] Bartłomiejczyk, M. Strategies of simultaneous interpreting and directionality[J]. *Interpreting*, 2006, 8 (1): 149 – 174.

[3] Brannen, J. *Mixed Methods research: A discussion paper*[M]. Southampton: ESRC National Center for Research Methods, 2005.

[4] Brenna, S. E. & M. Williams. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers[J]. *Journal of Memory and Language*, 1995, 34 (3): 383 – 398.

[5] Bühler, H. Linguistic (semantic) and extralinguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters[J]. *Multilingua*, 1986, 5 (4): 231 – 235.

[6] Chang, C. & D. Schallert. The impact of directionality on Chinese/English simultaneous interpreting[J].

Interpreting, 2007, 9 (2): 137 – 176.

[7] Cheung, A. The effectiveness of summary training in consecutive interpreting delivery[J]. *Forum*, 2007, 5 (2):1 – 23.

[8] Choi, J. 2013. Assessing the impact of text length on consecutive interpreting [C] // Tsagari, D. & R. van Deemter. *Assessment Issues in Language Translation and Interpreting*. Frankfurt am Main: Peter Lang, 2013. 85 – 96.

[9] Clifford, A. Discourse theory and performance-based assessment: Two tools for professional interpreting[J]. *Meta*, 2001, 46 (2): 365 – 378.

[10] Clifford, A. Putting the exam to the test: Psychometric validation and interpreter certification[J]. *Interpreting*, 2005, 7 (1): 97 – 131.

[11] Crystal, D. *A Dictionary of Linguistics and Phonetics*[M]. Malden/Oxford/Carlton: Blackwell, 2008.

[12] Darò, V., Lambert, S. & F. Fabbro. Conscious monitoring of attention during simultaneous interpretation[J]. *Interpreting*, 1996, 1 (1):101 – 124.

[13] Donovan, C. Teaching simultaneous interpretation into B: A challenge for responsible interpreter training[C] // Godijns, R. & M. Hinderdael. *Directionality in Interpreting: The 'retour' or the Native?*. Gent: Communication & Cognition, 2005. 147 – 166.

[14] Fulcher, G. *Testing Second Language Speaking*[M]. London: Longman/ Pearson Education, 2003.

[15] Gile, D. *Basic Concepts and Models for Interpreter and Translator Training*[M] (Revised). Amsterdam: John Benjamins, 2009.

[16] Green, A. & R. Hawkey. Marking assessments: Rating scales and rubrics[C] // Commbe, C., Davidson, P., O'Sullivan, B. & S. Storynoff. *The Cambridge Guide to Second Language Assessment*. New York: Cambridge University Press, 2010. 299 – 306.

[17] Han, C. Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach[J]. *Interpreting*, 2015, 17 (2): 255 – 283.

[18] Jones, R. *Conference Interpreting Explained*[M]. Manchester: St. Jerome Publishing, 1998.

[19] Knoch, U. Diagnostic assessment of writing: A comparison of two rating scales[J]. *Language Testing*, 2009, 26 (2): 275 – 304.

[20] Kurz, I. Conference interpreting: Quality in the ears of the user[J]. *Meta*, 2002, 46 (2): 394 – 409.

[21] Kurz, I. Conference interpretation: Expectations of different user groups [C] // Pöchhacker, F. & M. Shlesinger. *Interpreting Studies Reader*. London & New York: Routledge, 2002. 313 – 324.

[22] Lee, J. Rating scales for interpreting performance assessment[J]. *The Interpreter and Translator Trainer*, 2008, 2 (2): 165 – 184.

[23] Lee, S-B. Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances [J]. *Interpreting*, 2015, 17 (2): 226 – 254.

[24] Lee, Y-H. Comparison of error frequency in simultaneous interpretation A to B and B to A (Korean-English) [D]. Unpublished DEA (pre-doctoral) thesis, University of Geneva, 2003.

[25] Linacre, J. M. *A User's Guide to Facets: Rasch-model Computer Programs*. [Computer software and manual]. Retrieved April 10, 2005, from www. winsteps. com.

[26] Liu, M. Assessment[C] // Pöchhacker, F. *Routledge Encyclopedia of Interpreting Studies*. London: Routledge, 2015. 20 – 23.

[27] Ng, B. C. End users' subjective reaction to the performance of student interpreters[J]. *The Interpreters' Newsletter special issue*, 1992 (1): 35 – 41.

[28] Pöchhacker, F. Quality assessment in conference and community interpreting[J]. *Meta*, 2001, 46 (2): 410 – 425.

[29] Pöchhacker, F. Interpreting quality: global professional standards? [C] // Ren, W. *Interpreting in the Age of Globalization: Proceedings of the 8th National Conference and International Forum on Interpreting*. Beijing: Foreign Language Teaching and Research Press, 2012. 305 – 318.
[30] Pokorn, K. In defense of fuzziness[J]. *Target*, 2007, 19 (2): 327 – 336.
[31] Riccardi, A. Evaluation in interpretation: Macrocriteria and microcriteria[C] // Han, E. *Teaching Translation and Interpreting* 4: *Building Bridges*. Amsterdam: John Benjamins, 2002. 115 – 126.
[32] Roberts, R. P. Interpreter assessment tools for different settings[C] // Roberts, R., Carr, S. E., Abraham, D. & A. Dufour. *Critical Link* 2: *Interpreters in the Community*. Amsterdam: John Benjamins, 2000. 103 – 130.
[33] Roy, C. B. Response to Etilvia Arjona on curriculum design[C] // McIntire, M. L. *New Dialogues in Interpreter Education*. Silver Spring: RID, 1984. 36 – 42.
[34] Sawyer, D. B. *Fundamental Aspects of Interpreter Education: Curriculum and Assessment*[M]. Amsterdam: John Benjamins, 2004.
[35] Schjoldager, A. Assessment of simultaneous interpreting[C] // Dollerup, C. & V. Appel. *Teaching Translation and Interpreting* 3: *New Horizons*. Amsterdam: John Benjamins, 1995. 187 – 195.
[36] Setton, R. & A. Dawrant. *Conference Interpreting: A Trainer's Guide*[M]. Amsterdam: John Benjamins, 2016.
[37] Shohamy, E. *A Practical Handbook in Language Testing* [M]. Tel Aviv: Tel Aviv University, 1985.
[38] Skaaden, H. Assessing interpreter aptitude in a variety of languages[C] // Tsagai, D. & R. van Deemter. *Assessment Issues in Language Translation and Interpreting*. Frankfurt am Main: Peter Lang, 2013. 35 – 50.
[39] Tommola, J. & M. Heleva. Language direction and source text complexity: Effects on trainee performance in simultaneous interpreting[C] // Bowker, L., Cronin, M., Kenny, D. & J. Pearson. *Unity in Diversity: Current Trends in Translation Studies*. Manchester: St. Jerome, 1998. 177 – 186.
[40] Wang, M. W. & J. Stanley. Differential weighting: A review of methods and empirical studies[J]. *Review of Educational Research*, 1970, 40(5): 663 – 705.
[41] Weir, C. J. *Language Testing and Validation*[M]. Macmillan: Palgrave, 2005.
[42] Weigle, S. C. *Assessing Writing*[M]. Cambridge: Cambridge University Press, 2002.
[43] Wu, F. S. How do we assess students in the interpreting examinations? [C] // Tsagari, D. & R. van Deemter. *Assessment Issues in Language Translation and Interpreting*. Frankfurt am Main: Peter Lang, 2013. 15 – 33.
[44] 许宏晨. 第二语言研究中的统计案例分析 [M]. 北京:外语教学与研究出版社, 2013.

基金项目: 国家社科基金项目“口译学能测试任务模型构建实证研究”(21BYY064)

收稿日期: 2021 – 07 – 20

作者简介: 尚小奇,助理教授,博士,硕士生导师。研究方向:口译测试与评估。
李德超,副教授,博士,博士生导师。研究方向:翻译理论与翻译史。