对比研究英汉交替口译中量表式和意群式评分方式的评分者间信度

王怡安

指导教师：赵肖 助理教授

厦门 大学



学校编码：10384　　分类号_____密级 _____

学号：12020141152734　　UDC_____

厦门大学

硕 士 学 位 论 文

# 对比研究英汉交替口译中量表式和意群式评分方式的评分者间信度

## A Comparison of Inter-rater Reliability between Scale-based Rating and Proposition-based Rating in English-Chinese Consecutive Interpreting

指导教师姓名：　赵肖　助理教授

专 业 名 称：翻译硕士英语口译

论文提交日期：２０１７年　４月

论文答辩时间：２０１７年　５月

学位授予日期：２０１７年　６月

答辩委员会主席：_____

评　　阅　　人：_____

2017 年　月

# 厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

# 厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　月　日

# 摘　要

口译质量评估研究的一个中心问题是使用什么样的工具和标准（Sang-Bing Lee, 2015）。由于口译测试中评分者存在差异性以及口译资格考试具有高利害性，确保质量评估中的评分者间信度至关重要。从评估机构的角度来说，使用有效、可信的评估工具有助于保证评估结果的客观公正，对推动口译测试的标准化进程和口译人才的选拔都有重要意义。本文研究的是针对英汉口译质量评估中“信息准确度”这一标准采用两种评分方式对评分者间信度的影响以及背后可能的原因，旨在探讨口译评估中对评估方式的合理选择，保持评估结果的客观、公正。本研究选取厦门大学三级交替口译考试音频为样本，邀请 20 名评分者分别使用量表式（scale-based rating）和意群式（proposition-based rating）的评分表针对译文的信息准确度打分。评分者内部分为两组，一组为经验丰富的业内人士，另一组为口译专业三年级硕士研究生。使用 SPSS 工具分析整组和两组内部的分数一致性。通过对比使用两种不同的评分方式体现的评分者间信度（inter-rater reliability）的差异并结合访谈分析产生差异的原因。研究发现，两张评分表都体现了比较高的评分者间信度，意群式评分表的评分者间信度要高于量表式评分表。但是不同背景的评分者可能适用于不同的评分方式。对于经验丰富的职业译员来说，量表式或者意群式的评分方式体现的评分者间信度差别不大，但是具体操作时，意群式的评估表可以更有效地帮助评分者保持评估标准的一致性、客观性。而对于学生译员来说，使用意群式的评分方式可以帮助其在面对不同水平的口译表现时，

保持评分标准上的一致性，达到较高的评分者间信度。量表式评分表和意群式评分表各有特点，在实际操作中可以根据评分者的特点，评估的目的、性质等选择合适的方式。

**关键词：**口译质量评估　评分者间信度　信息准确度　量表式　意群式

# Abstract

Central to interpreting performance assessment is the choice of the instrument(s) and criteria to apply（Sang-Bing Lee, 2015）. Raters of a performance test vary and usually interpreting accreditation examinations are of high stakes, thus it is important to ensure the inter-rater reliability of the raters involved. From the perspective of the assessment agency, the use of effective and credible assessment tools can help ensure the objectivity of the assessment results, which is of great significance to the standardization of interpreting performance test and the selection of interpreting talents. This paper explores the difference of the inter-rater reliability when using two kinds of ways to assess the information accuracy of the test-takers' interpretation and the possible reasons behind the differences observed. In the experiment, the author selected samples of different levels from the English Interpreting Certificate test of Xiamen University and 20 raters were invited to score the interpretation samples in terms of information accuracy by two rating methods—scale-based rating and proposition-based rating respectively. Two groups of raters—experienced professional interpreters and novice interpreters (third-year graduate students majoring in English interpreting)—rated six consecutive interpreting performances with two rating methods and provided their feedback. The SPSS tool was used to analyze the inter-rater reliability of the whole group and the two sub-groups. The results showed that for the whole group, the inter-rater reliability of proposition-based

rating was a little higher than that of scale-based rating. For the experienced professional raters, there is little difference in raters' consistency of scoring between the two methods. But for student raters, it is obvious that proposition-based rating could better help them maintain the scoring consistency when dealing with samples of different levels—the inter-rater reliability of proposition-based rating was much higher than that of scale-based rating. Based on the analysis of the experiment results and the raters' feedback, the author found that both proposition-based rating and scale-based rating had their own advantages and may apply to raters of different backgrounds. Choosing the proper assessing method according to the purpose, raters, and nature of the interpreting performance test would help ensure the validity and reliability of the assessment.

**Key words**: Interpreting Performance Assessment; Inter-rater Reliability; Accuracy; Scale-based Rating; Proposition-based Rating

# 目 录

# Table of Contents

# 第一章　导言

## 1.1　研究背景

口译质量评估在口译教学、研究和资格认证中比较常见。口译质量测试作为重要的口译质量评估手段，应用范围非常广泛，包括口译专业入学考试、口译教学中的阶段测试和教学反馈、口译职业资格认证、口译产品质量评估等等。

在口译研究领域，虽然关于口译测试以及口译质量评估的研究比较广泛，但是针对口译质量评估本身的信度及效度的研究比较少，这反应出口译评估还主要依靠于主观理解而缺乏理论支持或者实证性研究（Pöchhacker, 2004; Sawyer, 2004）。口译质量评估与一般的客观性评估不同，并不是对正确答案或错误答案（如在选择题或判断题中）的简单计算，在评估的过程中不免要涉及评估人员的主观判断。如果在测试中使用的评估工具不能有效地帮助评分者达成一致的、清晰的判断标准，那么就可能对评估结果造成影响，也无法保障该测试本身的信度和效度。

影响口译质量本身的因素有很多。在测试环境下，测试本身，包括测试机构，输入形式，语速，文本类型，技术难度，话题新颖度和准备时间，都会影响口译质量（Bachman, 1990; Brindley, 1998）。既然口译测试是对口译质量进行评价，从理论上来说，影响评估结果的只能是被试者的口译质量而非任何其他因素。为了得到有效的口译质量评估结果，应该在测试开发，操作和评估的各个环节控制影响口译能力的外部因素，这就需要测试的设计者注重测试手段的有效性和可操作性，也需要对测试本身进行一定的监管和评估。从这个角度来说，选择合适的评估工具有助于帮助评分者更客观、有效地作出评价。本文作者就针对口译质量评估中

常见的“信息准确度”这一标准使用两种评分方式进行评分，通过对两者的评分者间信度进行对比研究，旨在探讨是否两种方式都能有助于保障不同背景的评分者在评分过程中保持一致的标准，对同一位被试者的口译表现给出趋于一致的分数。若两者存在差异，造成差异的原因可能是什么。通过对不同评分表评分者间的信度的对比研究旨在为高利害性口译测试的评估工具选择提供一定借鉴意义。

## 1.2　研究内容

口译质量评估研究的一个中心问题是使用什么样的工具和标准（Sang-Bing Lee, 2015）。在评估口译质量时，针对具体的口译任务、测试目的等会用到许多不同的方法。从口译测试设计者的角度来说，应当确保选择或设计的评估工具能够有效帮助参与评估的评分者对此次测试的评价标准达成一致的认识，评估的结果应当体现出较高的评分者间信度。本文研究的是针对口译质量评估中“信息准确度”这一项常见、重要的评估标准采用量表式和意群式两种评分方式进行评估，并针对得到的数据分析不同评分方式对评分者间信度的影响以及背后可能的原因，旨在探讨口译评估中对口译评估方式的合理选择，以保持评估结果的客观、公正。

## 1.3　研究方法

本文采用实验与访谈结合的研究方法。作者选取厦门大学三级交替口译考试音频为样本（英译汉，300 字左右），邀请 20 名评分者分别使用量表式（scale-based rating）和意群式（proposition-based rating）的评分表针对译文的信息准确度打分。评分者内部分为两组，一组为 10 名经验丰富的业内人士（以下简称 PR，职业口译员或高校口译老师或两者兼有），另一组为英语口译专业三年级硕士研究生（以

下简称 SR，已完成口译专业课程学习，具备一定口译素养，但经验较少）。评分结束后对每一位评分者进行访谈。使用 SPSS 工具分析整组和 PR、SR 组内的分数一致性。对比使用两种不同的评分方式体现的评分者间信度（inter-rater reliability）的差异并结合访谈分析产生差异的原因。通过数据对比和分析发现，两张表都体现出了较好的测试信度和效度，整组的评分者间信度比较接近，说明两张表都能帮助评分者保持一致的评分标准，作出评估判断。但是在 PR 和 SR 两个分组内使用两张表评分的评分者间信度存在差异，不用评分方式可能适用于不同类型的评分人员。

## 1.4　研究意义

由于口译测试中评分者存在个体差异性以及口译资格考试具有高利害性，确保质量评估中的评分者间信度至关重要。信度，指的是评估标准的一致性。评分者间信度具体可以体现为不同的评分者对于同一份被试样本打出的分数的一致性。具有较高评分者间信度的口译质量评估应当给质量相当的口译表现打出近似的分数。考生个人的表现、评分者、口译内容和评估标准都会对最后的评估结果产生直接影响。从评估机构的角度来说，使用有效、可信的评估工具有助于保证评估结果的客观公正，对推动口译测试的标准化进程和口译人才的选拔都有重要意义。本文通过对比研究针对口译评估中“信息准确度”这一常见的评估标准使用的两种评分方式所体现的评分者间信度，分析在针对口译信息的评分过程中影响评分者作出评分的影响因素，以及有效帮助评分者保持客观、一致性评分标准的要素。

这有对口译评估工具的选择具有借鉴意义。鉴于国内鲜少有学者对比研究量表式评分表与意群式评分表的评分者间信度，本研究具有一定创新意义。

## 1.5　全文框架

全文共分为五章。第一章为导言，介绍研究背景、研究内容、研究方法、研究意义。第二章为口译质量评估研究介绍，包括按照不同的分类方法产生的不同评分方式的特点，并介绍了评估表的信度和效度这两个重要概念、两者的关系以及评分者间信度的衡量方式。第三章为口译质量评估中的“信息准确度”标准及其评估方式，介绍了口译质量评估标准概况，信息准确度这一评估标准的重要性以及在本研究中用到的两种针对信息准确度的评分方式。第四章为两种评分方式的评分者间信度对比研究，介绍了实验方法、实验过程、实验数据分析和访谈反馈的结果。第五章为总结，得出本实验的研究结论，并对未来进一步深入研究提出建议。

# 第二章　口译质量评估

## 2.1　口译质量评估研究

口译质量的评估相对笔译起步较晚，再加上口译具有即时性、交互性，所以其评估模式也与笔译大不相同。Claudia V. Angelelli（2009）和 Holly E. Jacobson（2009）在其著作《口笔译学习测试与评估》中总结了有关口译质量评估研究方面存在的真空地带。自 1953 年国际会议口译员协会（AIIC）成立以来，口译领域的研究就一直专注于会议口译（Claudia V. Angelelli 和 Holly E. Jacobson，2009）。早期的实证性研究都集中于心理学范畴，研究同声传译的认知过程(Pöchhacker, 2004)。此外，正如 Hsieh（2003）指出的那样，同声传译方面的理论发展一直是由强调忠实与准确的翻译实践推动的。

然而，很少有学者专门研究口译整体方面的评估，尤其是口译质量方面的评估，以及在实证研究的基础上使用有效、可靠的手段评估口译质量。Angelelli (2001)使用心理测验学设计了第一个有效、可靠的评估工具研究口译员在加拿大、墨西哥和美国的各种工作场合中扮演的角色，比如法庭、医院、商业会议、国际会议和学校。Sawyer（2004）在美国进行了一项针对研究生口笔译能力评估的案例研究，并开始了关于口笔译测试效度验证的政治和道德后果的讨论。Clifford (2005）基于会话理论开发了一项口译资格认证考试。他认为，对于口译质量评估方面的研究还需要更多更严谨的方法。应当用实证性的研究完善资格认证考试工具的设计。