

托福iBT口语考试的评分等级评估

——基于中国考生的数据

汪雯, 许丹, 钟舒婷

新东方国际教育培训事业部, 北京

收稿日期: 2025年1月14日; 录用日期: 2025年2月17日; 发布日期: 2025年2月28日

摘要

随着在北美接受教育的中国学生数量的增加, 评估其英语水平的必要性也逐渐凸显, 尤其是在像托福iBT这样的高利害测试中。虽然托福口语测试对准留学生们来说是一个挑战, 但他们持续较低的口语成绩引发了关于其根本原因的讨论, 包括用于评估他们表现的评分量表描述。本论文旨在评估这些评分描述, 特别是综合口语任务的描述, 在多大程度上准确反映了中国考生的语言特征。本研究使用实证数据, 重点关注托福综合口语测试中的第二题, 该题结合了阅读、听力和口语技能。本研究选取了四个参加托福预备课程的中国学生的口语答题录音进行分析, 这些录音均获得3分(满分4分)的评分, 分析依据为托福口语测试评分量表的三个子构念: 总体陈述、语言使用和话题发展。研究结果显示, 考生的语言特征与托福评分量表描述之间存在若干差异。在总体陈述方面, 研究发现学生回答的语言特征(如流利度、发音和语调)与相应的描述只有部分一致。尽管大多数回答是可以理解的, 但学生没有表现出期望的流利度。例如, 回答的流利度未完全达到描述中期望的“流畅性”, 这表明流利度和可理解性可能并不总是相关。这种不一致可能表明这两个构建应该分开评估, 因为它们可能反映了口语表现的不同方面。在语言使用方面, 四个学生回答的语法和词汇使用存在若干差异。虽然每个回答表现出一些词汇准确性, 但在语法准确性和句子复杂性方面存在显著问题。这些差异往往导致交流不完整或不准确。该子构建的描述提到“某些不准确或不精确的词汇或语法结构使用”, 但未能明确允许的误差程度, 给评分者留下了解读空间。这种不明确性可能导致评分者根据主观解释对回答进行评分的不一致性。话题发展的分析也反映了托福口语测试评分描述的一些问题。尽管回答总体上连贯, 但往往缺乏具体性和完整性, 未能完全满足任务要求。回答中出现一些断断续续或不连贯的想法, 使研究人员难以将学生回答的语言特征与评分量表中的描述相匹配。虽然话题发展描述的模糊性(如“不完整或不准确的信息”)的确能为评分提供一定便利性, 但可能无法反映学生真实的话题扩展能力。研究结果表明, 尽管托福iBT口语测试描述为评估考生回答提供了有价值的框架, 但描述与考生产生的语言特征之间存在不匹配现象。这些差异可能影响评分过程的准确性和公平性。该论文呼吁需要在这领域进行进一步的研究, 以确保对学生口语能力的评估更加公平和有效。

关键词

托福口语, 评分标准, 测试效度, 测试信度

Evaluating the Rating Scale of the TOEFL iBT Speaking Test

—Based on Data from Chinese Test Takers

Wen Wang, Dan Xu, Shuting Zhong

Global Education and Training Office, New Oriental Education & Technology Group, Beijing

Received: Jan. 14th, 2025; accepted: Feb. 17th, 2025; published: Feb. 28th, 2025

Abstract

The increasing number of Chinese students studying in North America has highlighted the need to evaluate their English proficiency, particularly in high-stakes language tests like the TOEFL iBT. While the TOEFL speaking test is often a challenge for these students, their consistently low speaking scores have prompted discussions about the underlying causes, including the rating scale descriptors used to assess their performances. This paper aims to assess the degree to which these descriptors, particularly for the integrated speaking tasks, accurately reflect the linguistic features produced by Chinese test takers. Using empirical data, this study focuses on Task 2 of the TOEFL speaking test, which combines reading, listening, and speaking skills. This study selected four Chinese students who participated in TOEFL preparatory courses to analyze their spoken answer recordings, all of which received a score of 3 out of 4. The analysis is grounded in a content analysis framework that examines the three sub-constructs of the TOEFL speaking test's rating scale: delivery, language use, and topic development. The findings of this study highlight several discrepancies between the linguistic features exhibited by the test-takers and the descriptors in the TOEFL rating scale. In the area of delivery, the study found that the linguistic features of the students' responses—such as fluency, pronunciation, and intonation—only partially aligned with the corresponding descriptors. While most responses were intelligible, the students did not exhibit the desired level of fluency. For instance, the fluency of the responses did not fully meet the descriptor's expectation of "some fluidity," suggesting that fluency and intelligibility may not always correlate. This misalignment may suggest that these two constructs should be assessed separately, as they may reflect different aspects of speaking performance. For language use, several inconsistencies were found in the students' use of grammar and vocabulary. While some lexical accuracy was evident in the students' speech, there were significant issues with grammatical accuracy and sentence complexity. These discrepancies often resulted in incomplete or imprecise communication. The descriptors for this sub-construct, which mention "some imprecise or inaccurate use of vocabulary or grammatical structures," failed to clarify the degree of permissible error, leaving room for interpretation by raters. This lack of clarity could result in inconsistencies in how responses are rated, depending on the rater's subjective interpretation of the language use. The analysis of topic development also revealed some issues with the TOEFL speaking test's descriptors. While the responses were generally coherent, they often lacked specificity and completeness, failing to fully address the task requirements. Instances of choppy or disjointed ideas were identified, which made it difficult for the researchers to match the linguistic features of the students' responses with the descriptors in the rating scale. The vagueness of the descriptors for topic development, such as "incomplete or inaccurate information," allowed for flexibility in rating but may not reflect the true extent of the students' performance. The findings suggest that while the TOEFL iBT speaking test descriptors provide a useful framework for evaluating test-taker performance, there are notable mismatches between the descriptors and the linguistic features produced by the test-takers. These discrepancies may influence the accuracy and fairness of the scoring process. The paper calls for

further research in this area to ensure a fairer and more valid assessment of students' speaking abilities.

Keywords

TOEFL Speaking, Rating Scale, Test Validity, Test Reliability

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着在北美接受教育的中国学生数量呈指数级增长, 衡量他们使用和理解英语的能力变得愈发重要[1]。基于互联网的托福 iBT 考试是一项全球标准化的水平测试, 用于衡量“母语非英语人士的交际语言能力”[2]。该考试成绩也被高等教育机构用作招生参考[2]。自该考试实施以来, 中国学生面临着前所未有的压力, 因为他们在口语部分往往得分较低, 而考试结果可能对他们的学业和职业产生重大影响[3]。尽管部分学生口语分数较低可归因于考生自身的语言缺陷, 但这并不一定排除其他潜在因素, 如评分系统的影响[4]。托福分数通常是通过将评分等级的描述与考生的语言特征相匹配来生成的, 这增加了评分过程的主观性[5]。基于上述背景, 本研究旨在通过中国考生的实证数据回答以下研究问题:

托福口语考试评分等级的描述在多大程度上与考生的语言特征相匹配?

考虑到对如此广泛的评分级别进行详细分析的可行性, 本文仅评估托福口语综合题评分标准中的 3 分所对应的语言能力描述(见附录 A), 因为这是大多数中国学生所处的水平。为了给这一评估提供足够的理论依据, 本文将在文献综述中讨论与评分等级相关的关键概念以及对所选托福评分等级描述进行详细分析。下一部分将介绍所采用的方法, 包括所用数据、分析工具等信息。最后, 通过分析四个样本回答, 呈现并讨论研究课题的结果。研究结果可为改进托福 iBT 考试和评估过程提供参考。

2. 文献综述

2.1. 托福口语考试评分标准

由于托福考试旨在衡量考生的整体交际语言能力, 通常选择整体评分等级来评定考试结果[6]。整体评分分四个等级描述语言质量, 包括总体解释以及对三个构念(总体陈述、语言使用和话题展开)的详细阐述(见附录 A)。根据三个构念的综合影响, 每个任务表现会被赋予 0 至 4 分的单分数。4 分的回答必须具备所有三个构念的描述特征, 而较低水平的回答必须至少具备两个层面的特征[7]。整体评分等级被认为是“一个细致的、过程不断发展的结果”[8]。其早期版本基于坚实的理论模型, 特别是与交际语言能力相关的模型, 但为了确定评分等级每个等级的描述, 经从非英语母语学生收集的实证数据确定和确认了语言表现的不同特征[9]-[11]。然而, 这些研究中使用的样本回答可能无法代表大多数目标考生, 因为未提供与这些非英语母语学生相关的具体信息。因此, 在将这些分析中出现的任何证据应用于当前评分等级时应谨慎, 使用从目标考生收集的数据来评估相关描述是至关重要的。

2.2. 中级水平语言能力描述

为了确定综合口语任务中级水平的语言表现特征, 本节将根据之前对托福口语考试和应用语言学的

研究结果, 分析与总体陈述、语言使用和话题展开相关的描述(见附录 A)。

2.2.1. 总体陈述

总体陈述包含四个构念: 流利度、发音、语调及可理解度[7]。这个水平的典型回答具有一定的流畅性和较好的可理解度。发音、语调和可理解度等构念被归在表达类别下是有道理的, 因为 Jenkins 认为发音和语调(即句子重音)是可理解度的重要特征[12]。然而, 将流利度和可理解度结合起来似乎存在问题, 因为发音和语调良好的回答不一定流利。正如 North and Schneider 所指出的, “如果各个方面应作为单独的参数处理, 就不应将它们置于一起衡量” [13]; 因此, 有必要单独研究样本回答的可理解度和流利度。在本研究中, 可理解度将主要通过样本回答中“非目标发音”(即难以理解的单词)和“非英语语调”的综合错误来评估, 以确定它们是否表明具有良好的可理解度[11]。本样本回答的流利度将通过语速来确定, 因为语速被视为流利度的一个显著特征和“口语表现评分的重要预计因素” [7]。

2.2.2. 语言使用

中级水平的语言使用分析主要涉及语法和词汇, 特别是其使用的准确性和复杂性。这个水平的回答特点是“存在词汇或语法结构的一些不精确或不准确使用以及所使用结构范围有所局限” [14]。该描述进一步说明, 这些局限可能影响整体流利度, 但不应严重影响信息的传达。虽然这表明在这个水平的回答中任何语言应用不应存在严重干扰信息传达的情况, 但这并不一定代表回答者的实际语言表现。此外, 虽然语言使用的描述似乎也衡量连贯性, 但这个构念将在以下部分讨论(见 2.2.3)。

鉴于上述信息, 语言使用分析可参考: 1) 确定每个 T 单位, 指一个独立从句及其所有从句的语法错误[15]; 2) 由两名评分员仔细检查实义词(不包括语法词和逻辑连接词)的准确性; 3) 测量每个 T 单位的从句比例, 这被认为是语法复杂性的最佳指标之一[11]; 4) 使用基于网络的 VocabProfiler 工具测量每个样本回答的 TTR 指标[16], 从而确定词汇的丰富度; 5) 确定严重干扰信息传达的区域。

2.2.3. 话题扩展

第三个评分标准是话题展开, 考察回答与题目任务的“相关性”和“连贯性”, 与语言使用的部分内容重叠[7] [17]。根据托福评分等级的描述, 中级水平的话题展开大多是连贯和相关的, 但具有“一些不完整、不准确、内容缺乏具体性以及思路展开的不连贯”的特点[14]。因此从话题展开的角度分析样本回答, 研究者可以关注这些特定方面: 1) 与输入任务要求的关键信息相比, 不完整、错误和缺失的信息; 2) 任何存在逻辑连接词使用不当的不连贯区域, 并由两名评分员评估。

基于上述分析, 托福口语考试评分等级中, 体现了“表达”描述中的两个不同构念以及“语言使用”和“话题展开”描述中的各一个构念; 然而, 不确定这些复杂情况在多大程度上影响了评分过程。因此, 评分等级的描述与考生的语言特征的匹配程度还需要进一步探索。

3. 研究方法

3.1. 口语任务

由于本次评估针对综合口语评分标准, 选择了一个阅读/听力/口语任务(即 Task 2)。该任务要求考生在阅读一篇短文(例如: 一所大学取消诗歌写作课的公告)后, 听一段与主题相关的对话(例如: 讨论对该公告的意见), 然后给出一分钟的回答。

3.2. 口语样本

选择了四名中国学生的口语样本进行分析。这些学生此前参加过研究者教授的课程。所有样本都是针对 3.1 中描述的同一任务做出的回答, 并且根据学生的托福口语最终成绩预测会得到 3 分。这些样本

是在模拟考试环境中录制的,参与录制的学生正在准备即将到来的托福口语考试,因此代表了目标考生。从相关学生那里收集了签署的同意书,以确保数据的合法使用。

3.3. 研究步骤

采用混合方法分析内容并获得结果,涉及定性和定量分析。选择这种方法是因为它适合提取“语言交流”的“生产者属性”[18]。按照 Krippendorff 关于内容分析的指导方针,首先对数据进行转录,然后根据从评分等级分析(上文详细说明)开发的精心设计的框架(见附录 B)进行编码。为确保编码的可靠性,首先,由一名英语母语者使用相同框架对样本进行双重编码。其次,使用 Cohen 的卡帕系数检查编码间的可靠性[19]。基于编码,生成一组描述性统计数据来解释编码的语言特征并回答研究问题,同时提供本研究中出现的定性证据。

4. 研究结果与讨论

4.1. 总体陈述

如表 1 所示,发音和语调的综合不准确比例(表明可理解度)以及语速(表明流利度)在一定程度上与表达的相关描述相匹配。平均不准确比例(标准差 = 0.06 [0.02])表明样本回答中的大多数单词是可理解的,这与描述“讲话总体清晰,但在发音上有轻微困难”相符[14]。平均流利度(标准差 = 3.12 [0.38])表明四名同学都表现出相似程度的流利度。与 Munro 和 Derwing 提出的非母语者理想语速(4.7)相比,本样本回答的平均语速表明流利度较低[20]。然而,由于“一定”这个词的模糊性,这个特征不一定与描述“表达具有一定的流畅性”相匹配。

Table 1. Intelligibility and fluency of the sample responses (n = 4)

表 1. 示范回答中的可理解度和流利度(n = 4)

	可理解度		流利度(语速)		
	不准确数量/ 形符	比例	平均值(标准 差)	音节	平均值(标准 差)
R1 回答 1	10/138	0.07	0.06 (0.02)	193	3.2
R2 回答 2	11/110	0.1		158	2.6
R3 回答 3	8/139	0.03		212	3.5
R4 回答 4	9/146	0.05		195	3.2

另一个发现是,可理解度好(即发音和语调好)的语言表现不一定流利。例如,回答 2(R2)的语速为 2.6,是四个回答中最不流利的,可理解度的不准确比例为 0.1,这表明并非所有在这个水平得分的表现都同时展示出“良好的可理解度”和“一定的流畅性”。这进一步意味着流利度和可理解度是应该分别衡量的两个构念[13]。

4.2. 语言使用

4.2.1. 句子复杂性和词汇范围

如学生表达的从句的平均词汇范围(标准差 = 0.53 [0.02])和平均句子复杂性(标准差 = 1.67 [0.41])所示(见表 2),所有四个样本回答都显示出语言结构范围有限,因此与评分等级中的相应描述相匹配。然而,这种限制并非这个水平回答所独有,因为 Brown 等人研究中的较高水平也显示出类似的词汇范围(TTR = 0.56) [11]。这一发现也与 Douglas 的研究一致,该研究强调总体得分较高的考生在词汇使用上不一定优于得分较低的考生[21]。

Table 2. Sentence complexity and lexical range of the sample responses (n = 4)**表 2.** 示范回答中的句子复杂性和词汇范围(n = 4)

	句子复杂性	比例	平均值(标准差)	词汇范围	比例	平均值(标准差)
R1	17/10	1.3	1.67(0.41)	78/138	0.57	0.53(0.02)
R2	14/10	1.4		58/110	0.53	
R3	20/9	2.2		70/139	0.5	
R4	20/11	1.8		80/146	0.55	

4.2.2. 语法和词汇准确性

尽管在句子复杂性和词汇范围方面, 评分等级描述与样本回答之间存在明显的一致性, 但也发现了三个不一致之处(见表 3)。

Table 3. Task grammatical and lexical accuracy of the sample responses (n = 4)**表 3.** 示范回答中的语法和词汇准确性(n = 4)

	语法准确性	平均值(标准差)	词汇准确性	平均值(标准差)
R1	8/14(0.57)	0.6 (0.12)	44/47(0.94)	0.94 (0.03)
R2	6/9(0.66)		37/38(0.97)	
R3	4/9(0.44)		64/70(0.91)	
R4	8/11(0.73)		68/71(0.96)	

第一个不一致之处与词汇和语法使用的准确性有关。样本回答中较高的平均词汇准确性比例(标准差 = 0.94 [0.03])表明“词汇使用有效”; 然而, 由于平均语法准确性(标准差 = 0.6 [0.12])表明语法使用不太有效, 因此很难得出所有这些回答都展示出评分等级中描述的“词汇和语法的自动有效使用”的结论[14]。显然, 这一描述没有反映出那些词汇选择较好但语法使用水平相对较低的考生的真实表现[5]。

第二个不一致之处是很难将某些描述与数据中的语言特征联系起来。语法准确性比例存在一些显著差异, 最高为 0.73, 最低为 0.44。这种差异可能是由描述中的模糊措辞“词汇或语法结构的一些不精确或不准确使用”引起的[14]。因为这个描述没有明确说明在这个水平层面可允许的误差程度, 所以它创造了一个模糊空间[17], 使得评分者可以轻易地对回答表现进行评分。然而, 由于这个模糊的描述也给样本回答的语言特征匹配带来了困难, 评分者的这种做法是否对所有学生公平还有待观察。

在样本回答中还发现了与“这种限制不会严重干扰信息的传达”相关描述的不一致之处[14]。从以下样本回答的摘录中可以看出, 语法和词汇使用中的不准确之处阻碍了预期意思的清晰表达, 这表明评分等级的描述过于模糊:

“He think to take a poetry writing course in a nearby state university will make inconvenient is inconvenient for students.”
(R3)

“So they should and the bus the transportation is not very close to their university.” (R1)

这些发现佐证了 Fulcher 的假设, 即评分等级仅代表了评分者对考生表现的想象, 而不是他们的真实表现[22]。

4.2.3. 话题展开

通过分别对任何不完整或不准确信息和/或任何逻辑连接词的错误使用进行编码, 对所有四个样本回答的任务完成情况和连贯性进行了检查, 结果更有意义。如表 4 所示, 样本回答中的几个部分要么不完整、不准确, 要么过于笼统, 从而与评分等级中的描述“内容不完整、不准确、缺乏具体性”相匹配[14]。

同样，样本回答在“连贯性”方面，显示出“思路进展不连贯”，如下列摘录所示：

“And the second reason is that he think to take a poetry writing course in a nearby state university will make inconvenient is inconvenient for students because that will take more than one hours bus travelling to go to the other university to take that poetry. And it is not worthwhile for them to take.” (R3)

“For one thing, he thinks that in this small size class, he can have more occasion from professor and get better feedback from him. Also he can spend more time with professor.” (R4)

同样，“一定”这个词的模糊含义使得样本回答的语言特征与评分等级的描述难以准确匹配。

Table 4. Task fulfillment and coherence of the sample responses (n = 4)

表 4. 示范回答中的内容完整度和连贯性(n = 4)

	错误/不完整信息数量	不连贯数量
R1	3	1
R2	3	2
R3	4	3
R4	2	2

从上述发现和分析可以得出结论，四个中级水平样本回答的语言特征在一定程度上与托福 iBT 口语考试评分等级的相应描述只能达到一部分的匹配，这支持了 North 和 Schneider 的观点，即“评分等级不能保证有效性、准确性或平衡性”[13]。主要问题与几乎所有三个构念(即总体陈述、语言使用和话题展开)的描述中模糊的措辞(即“一定”)以及语言使用的实际表现与其相应描述之间的不一致有关。这些问题可能导致后果，因为它们评价考生的语言表现时鼓励了主观性。如果评分者遇到考生回答的一个特征与特定水平的描述不匹配或难以匹配，评分者则不得不依赖自己的印象或经验[17]。虽然评分者通常因不可靠的评分而受到指责，正如 North 所认为的，评分者可能不会遵循评分等级中规定的所有标准，但有证据表明评分等级中的措辞模糊实则鼓励了评分的主观性[23]。正如 Upshur 和 Turner 所述，“由于评分等级描述不够精确，评分者的标准在一次评分过程中可能会发生变化”[5]。此外，评分的主观性可能会威胁到分配给这些样本回答的分数的可靠性，因为正如中级水平评分等级所述，没有一个被评为 3 分的回答能够完全表现出三个构念中的至少两个。根据 Douglas 的说法，这种主观性反之会通过给出的考试分数影响对语言能力或习得的解释[21]。

为了解决这些问题，引入分析性评分等级是一种解决方案。这种具有更详细描述评分等级可以让评分者更好地识别每个水平的表现，从而提高评分过程的客观性。此前的一项研究评估了使用分析性评分等级来衡量托福口语表现的可能性[17]。虽然结果不太可能应用于实际，但该研究的积极尝试不可否认。与其取代当前的评分等级，一个更权宜的解决方案是对其进行改进。例如，可以通过对考生数据的深入分析提供解释性表现来澄清整体评分等级中的几个描述(即包含“一定”这个词的描述)[24]。

5. 结论

本研究旨在评估托福 iBT 考试中使用的综合口语任务的评分等级。通过内容分析研究了托福备考班中四名中国考生的四个样本回答(3 分)，结果显示这些回答中表现出的语言特征部分与评分等级的相应描述相匹配。调查还发现了评分等级描述中的一些不一致之处。然而，由于方法的局限性(如样本量小、编码者信度不高，见附录 C)，结论还有待持续验证。

评分等级中模糊或不一致的描述也可能影响结果，使得一些回答无法与其相应描述匹配。尽管存在这些局限性，但本项研究为后续研究提供了方向，例如验证是否可以采用数据驱动的方法来改进当前的

评分等级。

致 谢

首先,我们要向新东方北京国际教育美高英高部总监高磊老师和助理总监孟醒老师表示衷心的感谢。在他们的支持下,我们才得以顺利开展联合创作和发表工作。同时,我们也要感谢托福口语教研组长汪思蒙老师、期刊编辑老师和教学管理部刘晨宇老师,她们的专业校对和支持我们的论文更加完善。最后,我们还要特别感谢 Glenn Fulcher 教授在研究设计阶段提供的宝贵意见,老师的建议提升了本项研究的严谨性。文中如有任何疏漏或不足,均由我们作者自行承担。

参考文献

- [1] Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., *et al.* (2017) Preparing for the Speaking Tasks of the *TOEFL iBT®* Test: An Investigation of the Journeys of Chinese Test Takers. *ETS Research Report Series*, **2017**, 1-59. <https://doi.org/10.1002/ets2.12145>
- [2] Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P. and Taylor, C. (2000) TOEFL 2000 Framework: A Working Paper. (TOEFL Monograph Series Report No. 16). ETS Research Report Series.
- [3] ETS (2016) Test and Score Data Summary 2023. <https://www.ets.org/pdfs/toefl/toefl-ibt-test-score-data-summary-2023.pdf>
- [4] Brown, H.D. and Abeywickrama, P. (2010) *Language Assessment: Principles and Classroom Practices*. 2nd Edition, Pearson Education.
- [5] Upshur, J.A. and Turner, C.E. (1995) Constructing Rating Scales for Second Language Tests. *ELT Journal*, **49**, 3-12. <https://doi.org/10.1093/elt/49.1.3>
- [6] Weir, C.J. (1990) *Communicative Language Testing*. Prentice-Hall.
- [7] Jamieson, J. and Poonpon, K. (2013) Developing Analytic Rating Guides for TOEFL iBT's Integrated Speaking Tasks. *ETS Research Report Series*, **2013**, i-93. <https://doi.org/10.1002/j.2333-8504.2013.tb02320.x>
- [8] ETS (2010) TOEFL iBT Test Framework and Test Development. <https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v1.pdf>
- [9] Butler, F.A., Eignor, D., Jones, S., McNamara, T. and Suomi, B.K. (2000) TOEFL 2000 Speaking Framework: A Working Paper (TOEFL Monograph Series MS-20). ETS Research Report Series.
- [10] Enright, M., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissan, S., Powers, D. and Schedl, M. (2008) Prototyping New Assessment Tasks. In: Chapelle, C., Enright, M. and Jamieson, J., Eds., *Building a Validity Argument for the Test of English as a Foreign Language*, Routledge, 97-144.
- [11] Brown, A., Iwashita, N. and McNamara, T. (2005) An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks. *ETS Research Report Series*, **2005**, i-157. <https://doi.org/10.1002/j.2333-8504.2005.tb01982.x>
- [12] Jenkins, J. (2009) *World Englishes: A Resource Book for Students*. 2nd Edition, Routledge.
- [13] North, B. and Schneider, G. (1998) Scaling Descriptors for Language Proficiency Scales. *Language Testing*, **15**, 217-262. <https://doi.org/10.1177/026553229801500204>
- [14] ETS (2004) Scoring Guides (Rubrics) for Speaking. ETS. https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf
- [15] Hunt, K.W. (1970) Syntactic Maturity in Schoolchildren and Adults. *Monographs of the Society for Research in Child Development*, **35**, iii-iv+1-67. <https://doi.org/10.2307/1165818>
- [16] Cobb, T. and the Université du Québec à Montréal (2002) The Web Vocabulary Profiler (1.0) [Software]. <https://www.lex tutor.ca/vp/eng/>
- [17] Xi, X. (2007) Evaluating Analytic Scoring for the TOEFL® Academic Speaking Test (TAST) for Operational Use. *Language Testing*, **24**, 251-286. <https://doi.org/10.1177/0265532207076365>
- [18] Krippendorff, K. (1989) Content analysis. In: Barnouw, E., Gerbner, G., Schramm, W., Worth, T.L. and Gross, L., Eds., *International Encyclopedia of Communication*, Vol. 1, Oxford University Press, 403-407.
- [19] Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37-46. <https://doi.org/10.1177/001316446002000104>
- [20] Munro, M.J. and Derwing, T.M. (1995) Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second

-
- Language Learners. *Language Learning*, **45**, 73-97. <https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
- [21] Douglas, D. (1994) Quantity and Quality in Speaking Test Performance. *Language Testing*, **11**, 125-144. <https://doi.org/10.1177/026553229401100203>
- [22] Fulcher, G. (1987) Tests of Oral Performance: The Need for Data-Based Criteria. *ELT Journal*, **41**, 287-291. <https://doi.org/10.1093/elt/41.4.287>
- [23] North, B. (2000). The Development of a Common Framework Scale of Language Proficiency. Peter Lang US. <https://doi.org/10.3726/978-1-4539-1059-7>
- [24] Fulcher, G. (1996) Does Thick Description Lead to Smart Tests? A Data-Based Approach to Rating Scale Construction. *Language Testing*, **13**, 208-238. <https://doi.org/10.1177/026553229601300205>

附录 A

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as the speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.
2	The response is connected to the task, though it may be missing some relevant information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following:	Speech is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. Speech may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level.	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well-connected or cohesive so that familiarity with the stimulus is necessary to follow what is being discussed.
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt).
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

附录 B

Category	Subcategory	Coding unit	Calculation
Delivery	Pronunciation and intonation inaccuracy ratios	Incomprehensible words	Divide the number of incomprehensible words by the total number of words
Language use	Grammatical accuracy ratio	Error free T- unit	Divide the number of error free T-units by the total number of T-units
	Lexical accuracy ratio	Inappropriate use of content words (i.e. nouns, verbs, adjectives, adverbs)	Divide the number of incorrect content words by the total number of content words
	Sentence complexity	Dependent clauses per T-unit	Divide the number of dependent clauses by the total number of T-units
	Vocabulary range	Type-token ratio (TTR)	VocabProfiler
	Areas that seriously interfere with the communication of the message		Frequency counts
Topic development	Task fulfilment	Wrong/incomplete/general information	Frequency counts
	Choppiness in progression of ideas	Incorrect use of logic connectives	Frequency counts

附录 C

Category	Kappa
Grammar complexity	0.96
Grammar accuracy	0.91
Lexical accuracy	0.77
Intelligibility	0.75
Task fulfilment	0.72
Coherence	0.60