

雅思口语AI评分模型迭代

——雅思口语四项评分维度分数与口语总分关系的研究

何彧扬, 姚宇西*

新东方国际教育培训事业部总部, 北京

收稿日期: 2025年4月17日; 录用日期: 2025年5月15日; 发布日期: 2025年5月31日

摘要

雅思口语AI模考及评分系统正在教学中扮演愈发重要的角色, 其有效提高了教师批改作业和对学生个性化反馈的效率, 因此提升AI模考系统的评分准确性是一个重要的研究方向。本文通过对新东方八个分中心雅思教师的实际考试口语总分与对应的四项评分维度分数进行数据建模, 得出了通过4项维度分数计算口语总分的加权模型公式。本文也呈现了部分统计数据, 为后续口语模考平台迭代与学生针对性提分练习提供了支持。

关键词

雅思, 雅思口语, 雅思口语评分维度分数, AI模考, 数据加权建模

Iterative Development of AI Scoring Models for IELTS Speaking

—A Study on the Relationship between Sub-Scores of Four Rating Demensions and Overall Speaking Score

Yuyang He, Yuxi Yao*

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

Received: Apr. 17th, 2025; accepted: May 15th, 2025; published: May 31st, 2025

Abstract

AI-powered IELTS speaking mock test and scoring system is playing an increasingly important role

*通讯作者。

文章引用: 何彧扬, 姚宇西. 雅思口语 AI 评分模型迭代[J]. 国外英语考试教学与研究, 2025, 7(2): 54-59.
DOI: 10.12677/oetpr.2025.72007

in language teaching and has effectively improved the educators' efficiency in grading assignments and giving personalized feedback to students. Consequently, enhancing the accuracy of the mock test system is a vital focus of further research. This paper analyzes the real test scores from IELTS teachers in eight New Oriental Sub-centers and establishes a weighted formula for calculating the overall IELTS Speaking band from its four score breakdowns. Other than that, there are statistical findings presented to support the potential improvements of the AI-powered mock test platform and focused practice strategies for students.

Keywords

IELTS, IELTS Speaking, Score Breakdowns, AI Mock Test, Data Weighted Modeling

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/>



1. 引言

作为全球范围内认可度较高的英语语言能力测试，雅思考试(全称为 International English Language Testing System，简称为 IELTS)评分标准在各个国家和地区均保持一致。考试共分为听力，口语，阅读，写作 4 个科目，满分为 9 分。根据雅思官方最新的数据显示，2023~2024 年度中国大陆的考生在雅思学术类考试的平均总分为 5.9 分，各个部分的具体分数如下：听力成绩 5.9 分，阅读成绩 6.3 分，写作成绩 5.8 分，口语成绩 5.5 分。可以看出，口语成绩不足是导致中国考生难以在雅思成绩上取得突破的重要原因(见图 1) [1]。

For the latest version of IELTS statistics, please visit: <https://ielts.org/researchers/our-research/test-statistics>

Academic mean performance by nationality						
Nationality	Statistic	Listening Band Score	Reading Band Score	Writing Band Score	Speaking Band Score	Overall result
Azerbaijan	MEAN	6.9	6.6	6.1	6.5	6.6
Bangladesh	MEAN	6.0	5.6	5.6	5.8	5.8
Cambodia	MEAN	6.2	5.8	5.8	6.1	6.0
China	MEAN	5.9	6.3	5.8	5.5	5.9
Colombia	MEAN	6.7	6.8	6.0	6.5	6.6
Egypt	MEAN	7.0	6.6	6.1	6.7	6.7
France	MEAN	7.1	7.1	6.2	6.7	6.9
Ghana	MEAN	6.7	6.0	6.2	7.0	6.6
Hong Kong	MEAN	7.0	6.8	6.2	6.4	6.7
India	MEAN	6.6	6.1	6.1	6.2	6.3
Indonesia	MEAN	7.0	6.9	6.1	6.4	6.7
Iran, Islamic Republic of	MEAN	6.6	6.4	6.0	6.5	6.4

Figure 1. IELTS test taker performance data 2023~2024 excerpt
图 1. 2023 年至 2024 年的最新雅思数据节选

此外，雅思口语考试共有 4 项评分维度，分别是流利性和连贯性(fluency and coherence)、词汇多样性 (lexical resource)、语法多样性及准确性(grammatical range and accuracy)和发音(pronunciation) [2]。当前提到的口语 AI 模考平台，能够在分析学生输入的音频后，自主判断出四项评分维度的对应得分(4 项小分)，并通过求平均数的方式算出学生的口语模考总分。此算法的优点是让学生同等重视 4 项评分维度，但存在一个隐患，将四项小分求平均以得出口语总分的计算方式并未得到官方认可，其只是推测结论。在实际考试中已经观察到了一定比例的 4 项小分与口语总分不符合平均关系的案例。因此通过收集实考口语成绩与对应 4 项小分并进行数据建模，以寻找他们间可能存在的统计学关系，是 AI 口语模考后续发展的首要任务。

2. 数据收集与初步分析

2.1. 数据收集

本次数据建模使用了来自新东方国际教育事业部北京、深圳、济南、西安、厦门、杭州、珠海、天津共八个分中心的 90 条数据与 18 条网络数据。每条数据均包含实考口语总分与 4 项小分, 共 5 列(见图 2)。其中, FC 代表“流利性和连贯性”, LR 代表词汇多样性, GRA 代表语法多样性及准确性, PR 代表发音。

口语总分	FC (流利度)	LR (词汇)	GRA (语法)	PR (发音)
6.5	6	6	7	7
7.5	7	8	8	7
8	8	8	8	8
8	8	8	8	8
9	9	9	9	9
7.5	7	8	8	8

Figure 2. Total and sub-score excerpts from the oral test collection table
图 2. 实考口语总分, 4 项小分收集表节选

2.2. 数据处理

在进行数据建模前, 本研究先对数据进行了初步分析, 增加了以下 5 个自变量。1) “小分平均”: 由 FC, LR, GRA, PR 求平均数得出; 2) “异常”: 若小分平均与口语总分的差值大于 0.25, 则填入“是”, 反之填入“否”; 3) “精准”: 若小分平均与口语总分完全一致, 填“是”, 其余填“否”。4) “偏差”: 小分平均减去口语总分。5) “偏差描述”: 若小分平均小于口语总分, 填入“偏高”, 若小分平均大于口语总分, 填入“偏低”, 若小分平均等于口语总分, 填入“相等”(见图 3)。

口语总分	FC (流利度)	LR (词汇)	GRA (语法)	PR (发音)	小分平均	异常	精准	偏差	偏差描述	备注
7.5	8	8	8	8	8	是	否	0.5	偏低	北京
6.5	7	7	7	7	7	是	否	0.5	偏低	北京
7	6	7	6	7	6.5	是	否	-0.5	偏高	北京
6.5	6	6	6	6	6	是	否	-0.5	偏高	厦门
6	6	7	6	7	6.5	是	否	0.5	偏低	杭州
6	7	7	6	6	6.5	是	否	0.5	偏低	杭州
7.5	8	8	7	9	8	是	否	0.5	偏低	深圳
6.5	7	7	7	7	7	是	否	0.5	偏低	网络数据
6	6	6	5	5	5.5	是	否	-0.5	偏高	网络数据
7	8	8	7	7	7.5	是	否	0.5	偏低	网络数据

Figure 3. Data preprocessing, including mean, anomaly, precision, deviation and deviation description
图 3. 数据预处理成果图, 包含小分平均, 异常, 精准, 偏差, 偏差描述

2.3. 偏差情况

在共计 108 组样本中, 共有 52 组口语总分不完全等于 4 项小分平均数, 其中以“偏低”为主, 共有 43 组(见图 4)。总体看来, “偏低”的概率为 39.8%, 偏高的概率为 8.3%, 相等的概率为 51.9%。从中可以看出, 当出现口语总分不完全等于 4 项小分的平均数时, 向下取整的概率更大。

基于 43 个“偏低”案例, 我们可以来观察在 4 项小分中, 是否有哪个小分出现了最多的“低于总分”的情况, 用于预测导致总分偏低的主要原因。经过计算可以看到, 在 43 个“偏低”案例中, 出现了 17 起 GRA 低于口语总分的情况, 因此我们可以做出预测: 语法表现不佳更容易导致口语总分偏低(见图 5)。

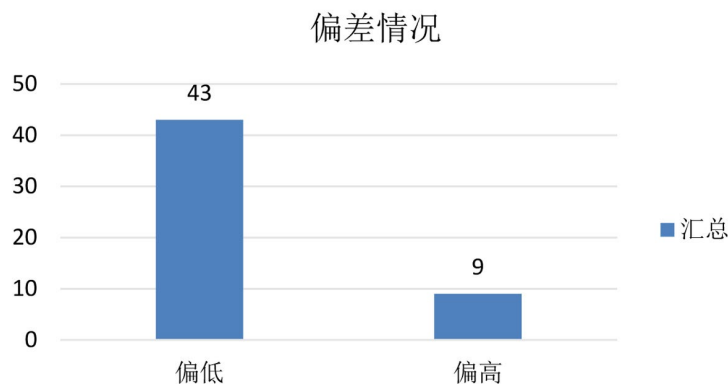


Figure 4. Ratio of deviations above/below the expected value

图 4. 在出现偏差时, 偏高与偏低的比例

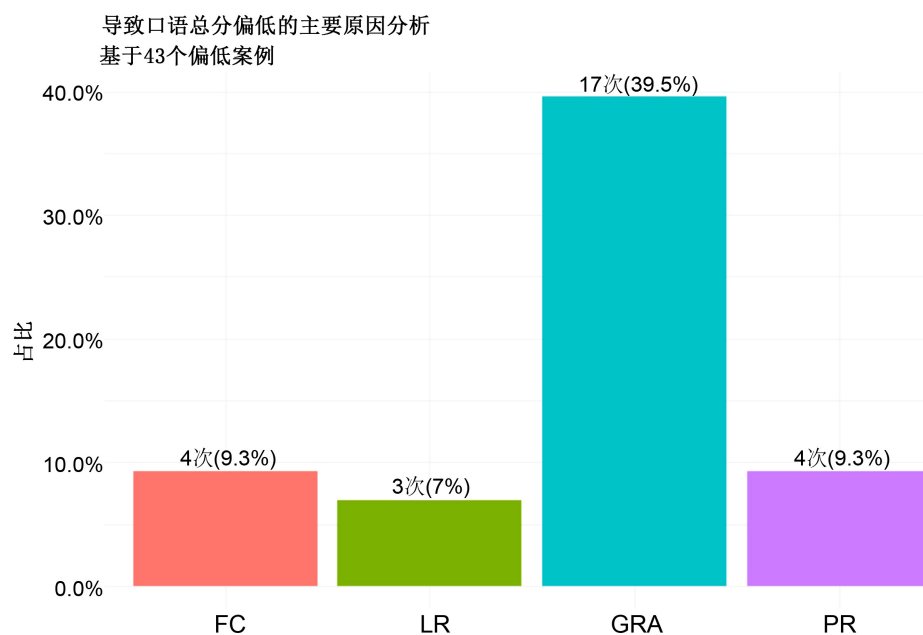


Figure 5. Key factors contributing to below-average oral performance

图 5. 导致口语总分偏低的主要原因分析

2.4. 异常情况

在本次收集的 108 组实考数据中, 有 10 组可被归类为“异常”的数据, 即 4 项小分的平均数与口语总分的净差值超过 0.25。其中“偏低”为主, 共 7 条, 即 4 项小分平均数低于口语总分, 包括一些不符合常理的分, 例如在 4 项小分均为 8 的情况下, 口语得分为 7.5。同时也存在“偏高”的情况, 一条来自厦门分中心的数据显示, 在 4 项小分均为 6 的情况下, 口语得分为 6.5。对于本次数据收集的总量而言, 10 条异常数据已经接近总量的 10%, 可以认定其并非人为操作失误或是突发情况。

当前并未发现异常情况的发生与考场地点或是口语总分有直接联系, 目前的猜测是其是否与雅思总分有联系? 但验证此猜想需要更多包含异常情况的数据作为支持。

3. 数据建模

根据上一部分做出的预测, 本次建模的主要目的是判断在 FC、LR、GRA、PR 这 4 项小分中, 是否

有哪一到两项对口语总分的影响更大? 因此, 模型因变量(dependent variable)为“口语总分”, 4 个自变量(independent variable)则分别为“FC”、“LR”、“GRA”、“PR”。模型选择为多元线性回归(multiple linear regression), 线性回归是研究变量间相互关系的一种有力工具, 是利用数理统计中回归分析, 来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法[3]。在 R-studio 环境下进行建模, 建模选用数据包括收集到的全部 108 个实考案例。

模型展示

经过 R-studio 环境建模, 结论为: $\text{总分} = 0.395 + 0.221 \cdot \text{FC} + 0.267 \cdot \text{GRA} + 0.240 \cdot \text{LR} + 0.207 \cdot \text{PR}$ 。4 个自变量的最右侧均有 3 个星标, 在 R-studio 建模环境下表示这四个变量都是显著变量(significant variables), 对模型均有重要影响, 同时本模型的拟合度为 91.42%, 不存在过拟合(over-fitting)的情况(见图 6)[3]。从各变量的系数可以看出 GRA 的系数最大为 0.267, 其次是 LR 的系数为 0.240。因此在 4 项小分出现相同幅度的变化时, GRA 对于口语总分的影响最大, 这也验证了前文对于“语法表现不佳更容易导致口语总分偏低”的猜想。

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39589    0.20139   1.966   0.052 .
FC           0.22058    0.03921   5.625 1.60e-07 ***
GRA          0.26681    0.03432   7.775 5.96e-12 ***
LR           0.23955    0.03930   6.095 1.92e-08 ***
PR           0.20675    0.03753   5.509 2.68e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1974 on 103 degrees of freedom
Multiple R-squared:  0.9157,    Adjusted R-squared:  0.9124

```

Figure 6. Screenshot of the R-studio model

图 6. R-studio 模型截图

4. 后续升级方案与对考生的启示

4.1. 模型后续改进方案

当前用于建模的源数据还可以进行两个方面的改进。其一, 当前的自变量仅包括雅思口语小分与总分, 并未参考雅思总分(包含听, 说, 读, 写)与考试地点等因素。在今后的数据收集可以加上这两个自变量再次建模并观察其对口语总分是否有显著影响。其二, 数据当前集中在有限的 7 个分中心且均来自教师, 也就是对考试有较深了解者。为更好地贴近学员/初学者的考试状态, 在今后的数据收集可以加入学生实考成绩与对应的 4 项评分维度分数。此外, 为了提升模型的准确性, 源数据量应超过 200 组, 因此, 建议在下一轮模型迭代时, 以包含学生、教师及 10 个以上分中心的 200 组数据作为起始标准。

4.2. 对考生的启示

当前的雅思口语教学当中, 对于前文提到的 4 个评分维度并没有明确的优先级区分, 且学生往往更在意表现力更强的流利度与发音。但根据本次建模情况来看, 语法与词汇量对于口语总分的影响或许更大, 因此在时间有限的情况下, 学生可以优先关注语法与词汇量方面的改进。

致 谢

本次数据收集&建模的研究离不开各位新东方老师们的大力支持。在此感谢姚宇西老师对我的指导

与点拨, 以及感谢北京、深圳、济南、西安、厦门、杭州、珠海、天津分中心每一位提供了宝贵实考数据的教师与教学管理者。

参考文献

- [1] 李旖林. 浅析雅思口语的教学难点及教学策略[J]. 国外英语考试教学与研究, 2025, 7(1): 1-6.
<https://doi.org/10.12677/oetpr.2025.71001>
- [2] 雅思口语评分标准[EB/OL].
<https://assets.cambridgeenglish.org/webinars/ielts-speaking-band-descriptors.pdf>, 2024-10-24.
- [3] 平雅露. 基于多元线性回归对建筑价格的影响分析[J]. 统计学与应用, 2020, 9(1): 1-3.
<https://doi.org/10.12677/SA.2020.91003>