Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2021, 3(3), 104-118
Published Online August 2021 in Hans. http://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2021.33012

Hans汉斯

关于影响托福写作自动化阅卷工具E-Rater评分因素的实证研究

李丰贤

广州新东方学校，广东 广州
Email: lifengxian5@xdf.cn

收稿日期：2021年8月6日；录用日期：2021年8月23日；发布日期：2021年8月30日

摘　要

本研究通过分析ETS和新东方合作的e-rater评分系统，通过控制变量和数据分析，探讨e-rater针对不同特征的作文评分情况，比如：1) 文章字数和文章得分的关联度；2) 文章分段和文章得分的关联度；3) 文章衔接词数量和文章得分的关联度；4) 文章错误数量和文章得分的关联度；5) 文章错误种类和文章得分的关联度；6) 文章平均词句长度和文章得分的关联度等等。

关键词

托福写作，自动化阅卷，电子评分，语料库，语言测试

Factors That Affect the Automated Scoring Engine E-Rater in the TOEFL Writing Section: An Empirical Study

Fengxian Li

Guangzhou New Oriental School, Guangzhou Guangdong
Email: lifengxian5@xdf.cn

Received: Aug. 6th, 2021; accepted: Aug. 23rd, 2021; published: Aug. 30th, 2021

文章引用: 李丰贤. 关于影响托福写作自动化阅卷工具 E-Rater 评分因素的实证研究[J]. 国外英语考试教学与研究, 2021, 3(3): 104-118. DOI: 10.12677/oetpr.2021.33012

Abstract

By analyzing the ETS e-rater scoring system which is now available on the New Oriental TPO website, and by analyzing the control variables and data, this study discusses the scoring performance of ETS e-rater mainly targeted at compositions with different characteristics, such as: 1) the correlation between the number of words and the essay scores; 2) the correlation between paragraph organization and the essay scores; 3) the correlation between the number of cohesive connectors and the essay scores; 4) the correlation between the number of errors and the essay scores; 5) the correlation between the types of error and the essay scores; 6) the correlation between the average sentence and word length and the essay scores.

Keywords

TOEFL Writing, Automated Scoring, E-Rating, Corpus, Language Testing

1. 托福写作部分介绍

托福考试是由 ETS (Educational Testing Service，美国教育考试服务中心)出题和举办的英语语言测试，主要考查本科生阶段学习所需要的英语听说读写四项语言能力。

在托福写作部分，考生需要完成两篇作文：综合写作(Integrated Writing)和独立写作(Independent Writing)。综合写作题目要求考生根据给出的阅读和听力材料，完成一篇基于学术材料的作文，主要考察学生的摘要、引用和信息整合能力。独立写作题目要求考生根据具体的写作话题和指令，对常见的话题(familiar topic)进行分析并且完成一篇议论文写作。时长方面，综合写作要求考生 20 分钟内容完成，独立写作则为 30 分钟。评分方面，每篇作文得分均为 0~5 分，考生最终作文部分的总分为两篇作文各自得分的平均分，精确到 0.5 分，然后根据换算机制换算为 30 分制。

考官在评分时，会针对考生的作文质量做出整体的评价(holistic scoring)。考官主要会关注考生是否能够在作文中搭建基本的文章逻辑框架，以及文章是否有合理使用例子和论证进行观点的支持和展开。考官还会根据考生作文的文章扣题程度、词汇和句式的多样性和准确性来表达观点[1]。

2. 托福写作 E-Rater 结构介绍

e-rater 在评分时，主要根据文章的多项维度进行打分，如 图 1所示，具体细则如下：

- 逻辑组织和展开论述(organization and development)；
- 积极特征(positive features)；
- 词汇复杂度(lexical complexity)；
- 词汇关联度(topic-specific vocabulary usage)；
- 语法错误(grammar)；
- 用法错误(usage)；
- 行文规则(mechanics)；
- 文章风格(style)；

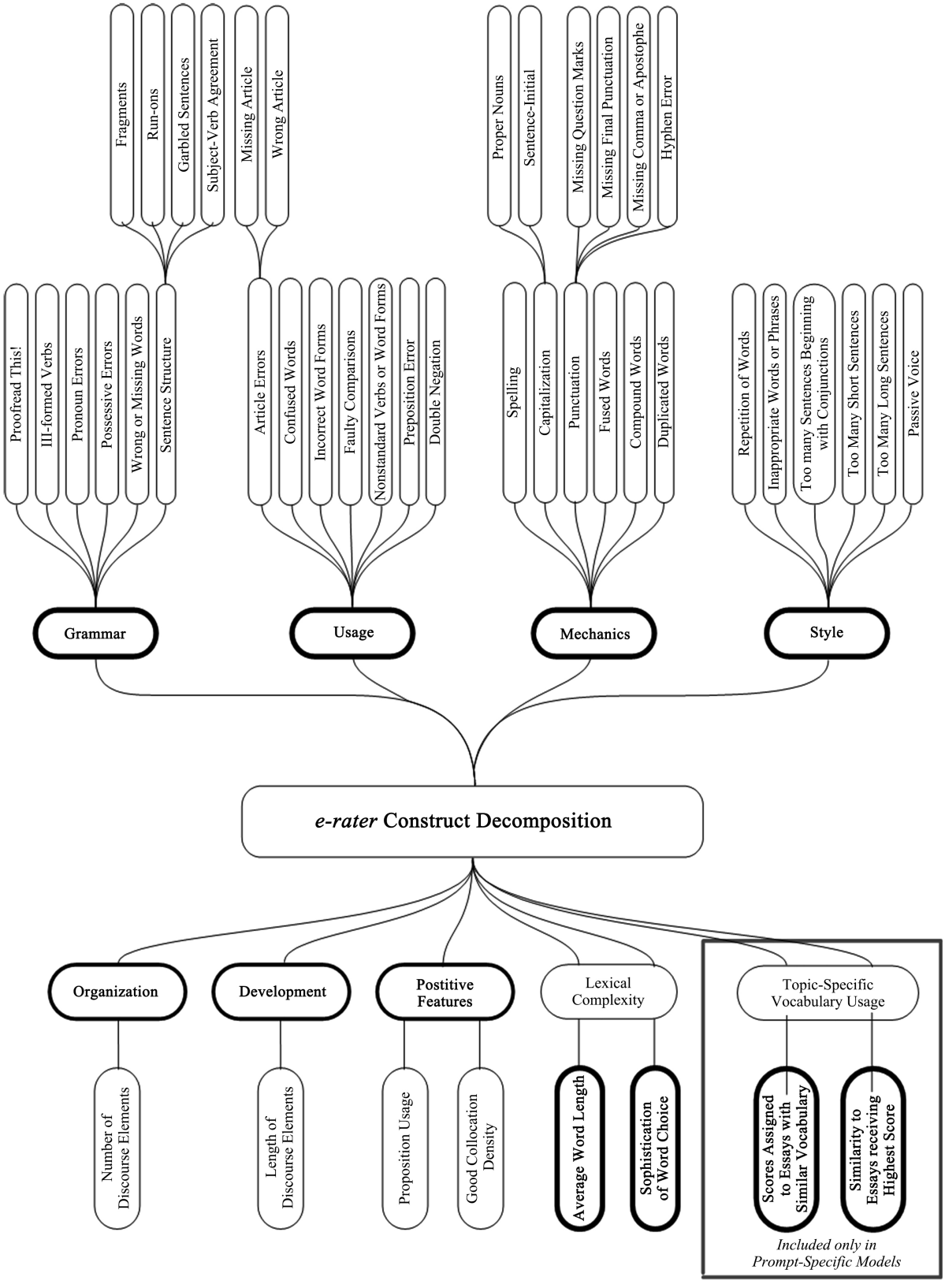

Figure 1. The e-rater construct decomposition [2]
图 1. e-rater 结构图[2]

3. 研究问题和研究过程

3.1. 本研究主要试图回答以下问题

1) 文章的字数是否会影响文章得分？
2) 文章的分段是否会影响文章得分？
3) 文章衔接词数量是否会影响文章得分？
4) 文章错误数量是否会影响文章得分？
5) 文章错误种类是否会影响文章得分？
6) 文章平均词句长度是否会影响文章得分？

3.2. 研究过程

本次研究主要选择托福 TPO28 综合写作[3]和 TPO30 独立写作[4]作为研究题目，具体题目请查看附录。本次研究的具体操作过程如下：笔者撰写一篇写作范文，经过 e-rater 评分后为 5 分满分的水平，然后再把这篇范文进行对应的修改和处理，每次控制 1~2 个变量，即对文章的修改，包括：文章字数、分段情况、衔接词数量、拼写错误、用词表达等等。然后再把修改后的文章重新放入到 e-rater 系统进行评分，最后进行相关得分和其他参数的记录，用作后续分析和解读。

4. 研究数据结果

4.1. 关于字数和文章的得分的关系

4.1.1. 具体参数

根据 表 1中 TPO28 综合写作作文编号 09、10 的情况来看，笔者主要修改是把满分 5 分范文中的听力细节删去，所以文章的字数随之下降(分别是 285 词、158 词)。对应地，e-rater 给文章的得分也有所下降，分别是 3 分和 2 分。

根据 表 2中 TPO30 独立写作作文编号 07、08、01 的情况来看，作文编号 07 删去了独立写作中所有的例子和原因，基本上只剩下大纲论点句和开头段，例子和原因的字数为 0；作文编号 08 则是有中等展开的例子和原因，字数为 161 词；作文编号 01 则是最原始的满分作文，有充分的例子和原因，例子和原因字数为 313 词。这三篇文章字数分别是 158 词、303 词、450 词，e-rater 分数分别是 3 分、4 分和 5 分。

根据 表 2中 TPO30 独立写作作文编号 09、10、11 的情况来看，作文编号 09 删去了 1 个主体段，剩余 2 个主体段，有充分的例子和原因，展开部分字数为 246 词；作文编号 10 删去了 2 个主体段，剩余 1 个主体段，剩余的主体段有充分的例子和原因，展开部分字数为 133 词；作文编号 11 删去了 3 个主体段，只剩下开头段和结尾段，展开部分字数为 0。这三篇文章字数分别是 359 词、223 词、61 词，e-rater 分数分别是 5 分、4 分和 1 分。

基于上述的情况，笔者还把 TPO30 独立写作的作文编号 01、08、09、10 进行了对比分析。作文编号 01 和 09 字数分别是 450、359，e-rater 给出的分数都是 5 分；而作文编号 08 和 10 号字数分别是 303、223，e-rater 给出的分数都是 4 分。

4.1.2. 具体参数分析

如 图 2所示，根据上述提及的这几篇作文情况来看，如果作文框架清晰并且没有各类语法和拼写错误，总体趋势是作文的字数越高，得分自然也越高。但是值得注意的是，当文章到了一定字数时，会存在一定的稳定性，比如作文字数从 223 提到 303 依然是 4 分，说明单纯增加字数不一定会带来分数的大幅提高，需要文章的各项参数都达到了一定的临界值才会让 e-rater 评为更高的分数档。

Table 1. The data results of TPO28 Integrated Writing samples
表 1. 关于 TPO28 综合写作样本的数据结果

编号	字数	句子总数	平均句子长度	福莱士阅读难度	错误数量	对文章的具体操作	e-rater 显示的主要错误种类	结构特点	e-rater 给分
1	304	13	23	49.5	20	随机选择 20 个单词末尾加字母 x，创造拼写错误	拼写错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
2	304	13	23	49	40	随机选择 40 个单词末尾加字母 x，创造拼写错误	拼写错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
3	304	13	23	47.3	60	随机选择 60 个单词末尾加字母 x，创造拼写错误	拼写错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
4	304	13	23	62.6	60	随机选择 60 个单词，用 x 替换，创造信息遗漏	冠词错误，拼写错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
5	244	12	20	59.3	60	随机选择 60 个单词，用空格替换，创造信息遗漏	句式残缺，单词遗漏	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4
6	302	29	10	62.9	0	把文章长句变改为短句	短句子过多	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
7	316	1	316	−245	0	把所有的句号改为 and 连接全文	粘连句，长句子	一句话写完整篇作文，大量使用 and 衔接，没有明确结构	0
8	285	13	22	50.7	0	把阅读信息全部写在一段听力信息写在另一个大段	无	阅读和听力各自为营，没有做到点对点反驳	5
9	189	8	24	48.6	0	把范文的听力细节删去	无	常规对比结构，有明确观点，但听力没有展开	3
10	158	8	20	52.4	0	把范文的听力观点和听力细节全部删去	无	常规对比结构，有阅读的明确观点，有听力模板，但没有听力的实质内容	2
11	304	13	23	47.9	60 + 16	随机选择 60 个单词末尾加字母 x，创造拼写错误加上 6 个主谓一致错误加上 10 个词性错误	拼写错误主谓一致词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5
12	302	13	23	47.3	60 + 26	随机选择 60 个单词末尾加字母 x，创造拼写错误加上 6 个主谓一致错误加上 10 个词性错误加入 10 个搭配错误	拼写错误主谓一致词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4

Continued

13	245	13	19	60.6	26	把 Sample 12 中的 60 个加上 x 的拼写错误删去，其他不变	主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4
14	302	13	23	47.3	300	把 Sample 12 中的剩余单词(除了衔接词外)全部都加 z，全部变成拼写错误	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	0
15	302	13	23	47.3	142	在 Sample 14 的基础上，进行改错，确保模板语言无错，同时修改其他错误，比如虚词 the、of、reading、listening、Peary、Tom	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	3
16	302	13	23	46.5	98	在 Sample 15 的基础上，进行改错，这次主要修复各类动词，包括谓语动词、非谓语动词等等	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4
17	302	13	23	47.3	250	在 Sample 14 的基础上，进行改错，随机减少错误，减少到 250 个单词	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4
18	302	13	23	47.3	200	在 Sample 17 的基础上，进行改错，随机减少错误，减少到 200 个单词	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	4
19	302	13	23	47.3	275	在 Sample 14 的基础上，进行改错，这次主要修复开头段，控制错误数量在 275 个	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	0
20	302	13	23	47.3	265	在 Sample 19 的基础上，进行改错，这次主要修复主体段的模板，也就是含有 reading 的字眼(但是不修复 listening)，控制错误数量在 265 个	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	0
21	302	13	23	47.3	250	在 Sample 20 的基础上，进行改错，这次主要修复主体段的模板，也就是含有 listening 的字眼，控制错误数量在 250 个	拼写错误 主谓一致 词性错误	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	3
22	324	43	8	66.3	0	把文章长句变改为短句	短句子过多	常规对比结构，有阅读的明确观点，有听力的有明确观点和听力细节展开	5

Table 2. The data results of TPO30 Independent Writing samples
表 2. 关于 TPO30 独立写作样本的数据结果

编号	对文章的修改	是否有背景句	是否有主论点	是否有开头段	是否有分论点	是否有主体段	是否有段落展开	主体段数量	展开句子数量	展开内容字数	文章总字数	衔接词数量	是否有结尾段	e-rater 给出分数
1	无	√	√	√	√	√	有，完全展开	3	11	313	450	15	√	5
2	删去结尾段	√	√	√	√	√	有，完全展开	3	11	313	429	15	×	5
3	删去背景句	×	√	√	√	√	有，完全展开	3	11	313	434	15	√	5
4	删去主论点	√	×	√	√	√	有，完全展开	3	11	313	426	14	√	5
5	删去开头段	×	×	×	√	√	有，完全展开	3	11	313	410	14	√	5
6	删去分论点	√	√	√	×	√	有，完全展开	3	11	313	353	12	√	5
7	删去全部展开例子和原因	√	√	√	√	√	无，没有展开	3	0	0	158	7	√	3
8	删去部分展开例子和原因	√	√	√	√	√	有，中等展开	3	6	161	303	10	√	4
9	删去一个主体段	√	√	√	√	√	有，完全展开	2	9	246	359	12	√	5
10	删去两个主体段	√	√	√	√	√	有，完全展开	1	5	133	223	10	√	4
11	删去全部主体段	√	√	√	×	×	无，没有展开	0	0	0	61	2	√	1
12	去除文章分段	√	√	全文只有一段	√	全文只有一段	有，完全展开	1	11	313	450	15	全文只有一段	5
13	去除文章分段 + 删去开头段结尾段内容	×	×	×	√	√	有，完全展开	3	11	313	389	12	×	0
14	删去所有衔接词	√	√	√	√	√	有，完全展开	3	11	282	414	0	√	5
15	去除文章分段 + 删去所有衔接词	√	√	√	√	√	有，完全展开	3	11	282	414	0	√	0
16	去除文章分段 + 删去所有衔接词 + 所有句号换为 and	全文只有一句	全文只有一句	全文只有一句	全文只有一句	√	有，完全展开	1	1	431	431	0	√	0
17	基于 16 的文本，保留一句话主论点单独成句	√	√	全文只有 2 句	全文只有 2 句	√	有，完全展开	1	1	392	430	0	√	0
18	根据 Sample 7 的内容把文段部分难词全部换成简单词	√	√	√	√	√	无，没有展开	3	0	0	158	7	√	2

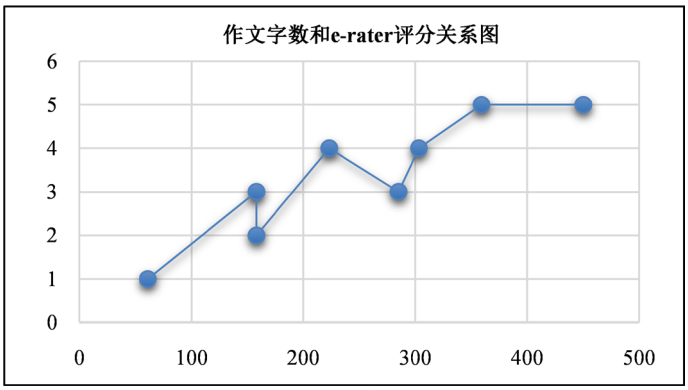

Figure 2. The correlation between the number of words and the essay scores given by the e-rater
图 2. 作文字数和 e-rater 评分关系图

4.2. 文章分段、衔接词数量和文章得分的关联度

4.2.1. 具体参数

根据 表 2中 TPO30 独立写作作文编号 01 至 06 的情况来看，这六篇作文的字数分别为 450、429、434、426、410、353，主要修改分别是删去结尾段、删去背景句、删去主论点、删去开头段、删去分论点，e-rater 对这六篇作文均给出了 5 分满分。

根据 表 2中 TPO30 独立写作作文编号 12 至 17 的情况来看，这六篇作文的字数分别为 450、389、414、414、431、430，主要修改分别是去除文章分段、去除文章分段以及删去开头段结尾段内容、删去所有衔接词、去除文章分段以及删去所有衔接词、去除文章分段以及删去所有衔接词以及所有句号换为 and、保留一句话主论点单独成句，e-rater 对这六篇作文打分分别是 5、0、5、0、0、0。

4.2.2. 具体参数分析

通过对上述这几篇作文情况分析，具体解读情况如下。首先，e-rater 非常看重文章的逻辑组织和展开(organization & development)，因为 e-rater 评分的起点就是根据文章的逻辑组织框架来读取并且分析文章。如果一篇文章没有清晰的逻辑框架组织，即使内容是满分，e-rater 也无法从文本读取如衔接词、观点句等的标志性元素，那么 e-rater 就会无法识别文章后续内容以及读不懂。在这种情况下，e-rater 就可能判低分或者零分。

相反，只要文章含有一定量的逻辑组织标记和标记，e-rater 就可以识别并且启动正常的评分程序，比如文章分段、使用衔接词等等。但是如果一篇文章既没有分段，也没有衔接词等等，这样的作文在 e-rater 当中就既有可能判为低分或者是零分，尽管文章的用词水平很高。根据 TPO30 独立写作的表格参数来看，虽然一篇文章不会因为衔接词数量的增多而大幅提高分数，但是如果文章缺少了衔接词，这会增大 e-rater 评分的难度，更容易触发 e-rater 零分或者低分的机制。

值得一提的是，本次研究尝试把 TPO28 综合写作的满分范文的所有句号都改为 and，去除文章的句号和句子切割，把整篇文章变成一篇超长的一句话作文，最后结果 e-rater 判为零分。从参数上来看，这篇作文触发了 e-rater 的零分机制。为了进一步调查，笔者使用了福莱士可读性指数(Flesch Readability) [5] 来作为文章的参考数据。一般来说，如果一篇英语文章的福莱士可读性指数介于 0~30，那么说明这篇文章的可读性较低且难度较大，对于大部分读者基本属于难以读懂。经过测量，这篇一句话作文的福莱士可读性指数是“−245”，那么从 e-rater 的角度来看来说就是无法读懂甚至是佶屈聱牙的作文，所以判为零分也不足为奇。

换言之，e-rater 对文章的组织和展开的设定其实反映了真人考官阅卷的过程：

1) 先从基本文章框架入手；

2) 识别主论点和分论点；

3) 评估主体段内容的数量和质量；

4) 检测文章扣分项，如各类语法错误和细节。

此外，e-rater 最看重的依然是文章主体段的展开例子和原因部分，即便文章删去了开头段和结尾段甚至是分论点，依然不影响文章可以拿 4 分高分或者 5 分满分。

4.3. 文章错误数量、文章错误种类和文章得分的关联度

4.3.1. 具体参数

根据 表 1中 TPO28 综合写作作文编号 01 至 03 的情况来看，三篇文章的字数没有变化，基本框架符合常规综合写作的逻辑组织和分段，错误数量呈递增趋势(分别是 20、40、60)，笔者通过随机选择单词加入字母 x 来制造拼写错误，文章的错误种类也基本集中在拼写错误。但是随着拼写错误的增加，e-rater 对这三篇作文的评分并没有改动，依然是 5 分满分。

根据 表 1中 TPO28 综合写作作文编号 11 至 13 的情况来看，这三篇作文基本框架符合常规综合写作的逻辑组织和分段。除了原有的 60 个拼写错误以外，作文编号 11 还分别增加了 6 个主谓一致错误、10 个词性错误；作文编号 12 则在此基础上增加了 10 个搭配错误；作文编号 13 则主要把文章的拼写错误修复去除，共剩下 26 个错误，错误主要类型为主谓一致错误、词性错误、搭配错误。三篇作文的错误数量为 76、86、26，e-rater 对这两篇作文的评分为 5 分、4 分和 4 分。

根据 表 1中 TPO28 综合写作作文编号 14 至 16 的情况来看，这三篇作文基本框架符合常规综合写作的逻辑组织和分段。错误数量分别是 300、142、98。作文编号 14 的作文种类则全是拼写错误；作文编号 15 和作文编号 16 的错误种类主要是拼写错误、主谓一致错误、词性错误。随着错误数量的减少，e-rater 给这三篇文章的打分呈上升趋势，分别是 0 分、3 分和 4 分。

根据 表 1中 TPO28 综合写作作文编号 17 至 18 的情况来看，这两篇作文基本框架符合常规综合写作的逻辑组织和分段。错误数量分别是 250 和 200。错误种类主要是拼写错误、主谓一致错误、词性错误。虽然错误数量有所下降，但是 e-rater 对这两篇作文的给分依然都是 4 分。

4.3.2. 具体参数分析

通过对上述这几篇作文情况分析，具体解读情况如下。首先，当作文的逻辑组织框架清晰和段落展开良好时，e-rater 对学生作文的拼写错误有较大的容错率。如 表 3所示，即便作文中含有 60 个拼写错误，但是整体文章的错误率(错误数量 ÷ 总字数)并没有特别高，所以没有对文章的参数造成太大影响，并且 e-rater 依然给出 5 分满分。这一点和 ETS 官方研究报告[6]的观点是一致的。

另一方面，拼写错误在错误种类当中对文章分数的贬损并不会特别严重。但值得注意的是，当拼写错误数量过高时，高出特定的参数和比率时，这会触发 e-rater 的零分机制，判定作文为非英语语言的作文(written in a foreign language)。

另一方面，普遍来说基础中等的学生在实际考场的作文并不会只有单一类别的拼写错误，一般都会含有多种类型的错误比如主谓一致等等。为了模拟学生在考场实际的作文情况，笔者故意在文章中加入其它类型的错误，具体参考 TPO28 综合写作的作文编号 11 和 12。这次实验再次证明：当作文逻辑组织框架清晰和段落展开良好时，即便文章有多种类型的语法和拼写错误时，e-rater 依然可以给出 5 分满分。但是随着错误种类和错误数量的增加，e-rater 则对文章进行扣分，从满分 5 分降至 4 分。从这项数据可以看出，当文章的特定参数达到一定临界值时，会触发 e-rater 的扣分机制。

Table 3. The correlation between errors and the essay scores of TPO28 Integrated Writing
表 3. TPO28 综合写作的作文错误和 e-rater 评分关系表

编号	字数	错误数量	错误数量/字数	e-rater 显示的主要错误种类	e-rater 给分
1	304	20	7%	拼写错误	5
2	304	40	13%	拼写错误	5
3	304	60	20%	拼写错误	5
4	304	60	20%	冠词错误，拼写错误	5
5	244	60	25%	句式残缺，单词遗漏	4
11	304	76	25%	拼写错误 主谓一致 词性错误	5
12	302	86	28%	拼写错误 主谓一致 词性错误	4
13	245	26	11%	主谓一致 词性错误	4
14	302	300	99%	拼写错误 主谓一致 词性错误	0
15	302	142	47%	拼写错误 主谓一致 词性错误	3
16	302	98	32%	拼写错误 主谓一致 词性错误	4
17	302	250	83%	拼写错误 主谓一致 词性错误	4
18	302	200	66%	拼写错误 主谓一致 词性错误	4
19	302	275	91%	拼写错误 主谓一致 词性错误	0
20	302	265	88%	拼写错误 主谓一致 词性错误	0
21	302	250	83%	拼写错误 主谓一致 词性错误	3

另外，笔者把 e-rater 判为零分的作文编号 14 进行错误修改，随着错误的减少，e-rater 评分也相应提高，这说明修改错误和减少文章的错误数量是有助于学生提高写作分数的，而且对于教师和学生而言，在不增加新内容的情况下，修改错误是学生提高写作得分较为高效的策略。毕竟，学生新增内容还可能会引入更多新的错误，可能会导致写多错多的情况。

类似地，根据作文编号 17 和 18，当作文总体质量达到一定层次比如 4 分的水平时，修改错误可能不会对分数有大幅度的影响。这和前面字数和分数的规律基本一致：当文章错误数量处于一定区间时，作文分数会存在一定的稳定性，这也说明在一些情况下仅仅减少错误不一定会带来分数的大幅提高，需要文章的各项参数都达到了一定的临界值才会让 e-rater 评为更高的分数档。

4.4. 文章平均词句长度和文章得分的关联度

4.4.1. 具体参数

根据 表 1中 TPO28 综合写作作文编号 01、06 和 22 的情况来看，这三篇作文基本框架符合常规综合写作的逻辑组织和分段，内容和 TPO28 综合写作的阅读和听力原文一致，文章的平均句长呈现下降趋势，分别是 23、10 和 8，整体文章无明显错误。数据结果显示，尽管文章平均句长逐渐减少，三篇文章的 e-rater 评分都是满分 5 分。

根据 表 4中 TPO30 独立写作的作文编号 07 和作文编号 18 的情况来看，这两篇作文基本框架符合常规独立写作的逻辑组织和分段，都没有充分的例子和原因，没有人为添加的错误。笔者的处理主要是把作文编号 07 的 20 个单词进行简化替换，改为更简单以及更短的单词，形成作文编号 18，具体单词对照表请查看附录 3。具体来说，两篇文章的福莱士可读性指数为 49.5 和 69.2，说明使用短单词的作文编号 18 比作文编号 07 要更容易读懂。词汇的平均长度方面，从字母数量来看分别是 4.8 和 4.2；按照音节数量来计算，作文编号 07 的多音节词占比为 21%，而作文编号 18 的多音节词占比为 9%。分数方面来看，e-rater 给这两篇作文判分为 3 分和 2 分。

Table 4. The vocabulary profiles of TPO30 Independent Writing sample No. 07 and No. 18
表 4. TPO30 独立写作作文编号 07 和 18 的词汇参数

编号	对文章的修改	福莱士阅读指数	平均词长(字母)	单音节词占比(%)	双音节词占比(%)	三音节词占比(%)	e-rater给出分数
7	删去全部展开例子和原因	49.5	4.8	62	17	21	3
18	根据 Sample 7 的内容把文段部分难词替换成更短音节的简单词	69.2	4.2	69	22	9	2

4.4.2. 具体参数分析

通过对上述这几篇作文情况分析，具体解读情况如下。首先，词汇和句式的长度和复杂度的确是 e-rater 考核的指标。但是作文平均句长的长短对作文总体得分影响不大，尤其是和文章的逻辑组织和展开论述这两个关键因素对比，平均句长这一指标显得比较相形见绌，不会有非常大的决定性影响。另一方面，平均词汇的长度对作文的影响则比较明显。在不改变原文框架以及也不增加新内容的前提下，把文章的平均词汇的长度提高能够一定程度上提高 e-rater 给文章的打分。

5. 对教学的启示和小结

关于作文的展开论述(development)方面，综合写作的主要得分要点都在听力部分，也就是说如果综合写作的作文要想进一步展开并且获得更高的分数，听力细节的数量是关键因素之一。对于独立写作来

说，举例和原因的占比也是决定文章分数的关键因素。教师可以重点设计课程和任务来训练学生段落拓展和展开观点的能力。

关于逻辑框架(organization)方面，针对基础薄弱的学生，教师可以把教学重点放在文章基本框架的搭建上，有了基本框架 e-rater 基本可以给出 3 分左右的分数。另一方面，从优先级来看，在段落框架中，教师应该把写作的教学重点放在主体段而不是开头段和结尾段，因为主体段是决定文章分数的核心。甚至教师之后不需要花太多精力让学生写开头段的背景铺垫，一句话表明主论点可能对学生而言更好操作。换言之，即便学生在真实考场写作没有写太多开头段和结尾段，只要主体段质量足够优秀并且展开充分，学生依旧可以获得写作的好分数。

关于衔接词(logical connector)方面，目前 e-rater 就是依靠分段和衔接词来标记文章的中各个板块，比如主论点、分论点、举例、展开、结论等等。但是没有了这些“路标”和清晰分段，e-rater 就会难以读懂，触发零分机制。有意思的是，当笔者想要进行词汇升级使用“from my perspective”表达论点时，e-rater 并没有识别出来。对比来看，笔者换成最为普遍的“in my opinion”后，e-rater 就成功识别了文章的主论点并且进行标记。由此可见，笔者建议教师在教授学生时使用常见的衔接词而不是标新立异的表达。根据本次实验，的确存在一种情况是不太常见的衔接词 e-rater 是无法读懂或者识别，因为 e-rater 只能读懂有限的衔接词。从这个维度来看，e-rater 可能不鼓励学生在衔接词这块的新颖性表达。

关于错误修改(errors)，针对基础薄弱的学生，在逻辑框架搭建好的前提下，可以优先针对高频错误进行修改，比如拼写错误、主谓一致错误、单复数错误等等。根据上述实验结果，笔者也建议托福写作教师在授课实践中，优先让学生把过往作文的错误修改打磨好，再开启新的作文写作任务。这样一来可以提高学生对自己高频错误形成特定的意识，同时也让学生获得持续的作文正向反馈，看到自己的作文只要修改错误就有机会获得更高的分数。

此外，笔者在实验过程中发现，对于 e-rater 而言，并不是作文中的所有错误都可以识别出来，一些比较细节或者无伤大雅的错误其实暂时先略过，比如标点错误、冠词错误等等。等到学生语法基础和各项能力进步了再回来纠正效果会更好。因为 e-rater 目前无法识别学生作文中的所有错误，所以学生的细节语言质量还是需要依靠写作老师批改作文和给反馈。

关于句子平均句长和平均词长(average sentence length and average word length)方面，在综合写作的评分中，作文的长句或者短句不一定会影响得分，关键是看句子表达的内容是否和原文一致。文章的一致程度越高，分数越高。长难句可以给基础较好的学生训练，起到锦上添花的作用，但核心得分点还是以内容为主。在独立写作方面，针对基础较好的学生，教师可以补充一些难度较大的学术词汇作为语料供学生选择，但是没有必要强迫学生必须使用高难度的词汇。因为即便是对于基础较好的学生，他们在使用学术词汇时，由于不习惯和陌生程度较高，也更有可能在使用的过程汇中犯错，比如拼写错误、搭配错误、词性错误等等。e-rater 的系统当中有一项加分项(Positive Feature)，也就是正确使用词汇搭配和介词。而这两块是很多学生的薄弱项。根据笔者过往的教学经验，中高分段的学生在介词和搭配这两部分也不一定会表现特别好，更不用说基础薄弱的中低分段学生。如果是这样，学生尝试通过升级词汇的方式来提分有可能会适得其反。

一言蔽之，本次研究发现 e-rater 评分的各项因素基本有效合理，托福写作教师可以根据上述研究结论优化课程，同时给学生正确方向引导，不断鼓励学生提高写作实力。只有写作硬实力到位，学生才可以产出无论是 e-rater 还是真人考官都认可的高分作文。

参考文献

[1] 杜璟, 冷楠. 对 GRE 写作自动化评分器 e-rater 评分准确性的实证研究[J]. 国外英语考试教学与研究, 2020, 2(3):

140-148. https://doi.org/10.12677/OETPR.2020.23013

[2] Quinlan, T., Higgins, D. and Wolff, S. (2009) Evaluating the Construct Coverage of the E-Rater® Scoring Engine. ETS, Princeton, NJ. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

[3] https://tpo.xdf.cn/question-analysis/write?quesNo=1&qId=23504&activeTab=video

[4] https://tpo.xdf.cn/question-analysis/write?quesNo=1&qId=23540&activeTab=video

[5] http://www.readabilityformulas.com/free-readability-formula-tests.php

[6] An Investigation of the *e-rater*® Automated Scoring Engine's Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model, ETS. https://files.eric.ed.gov/fulltext/EJ1168485.pdf

附录

附录 1. TPO28 综合写作阅读和听力原文

Reading

Robert E. Peary was a well-known adventurer and arctic explorer who in 1909 set out to reach the North Pole. When he returned from the expedition, he claimed to have reached the pole on April 7, 1909. This report made him into an international celebrity. Though some historians have expressed doubts that Peary did in fact reach the North Pole, three arguments provide strong support for the truth of Peary's claim.

First, the National Geographic Society put together a committee that was instructed to conduct a thorough investigation of Peary's records and equipment. At the end of the investigation, the committee concluded that Peary's accounts were consistent and persuasive and declared that he had indeed reached the North Pole.

Second, a recent expedition provides support for Peary's claim that he reached the North Pole in only 37 days after setting out from Ellesmere Island off the coast of Greenland. Skeptics used to argue that Peary could not have traveled that fast, since even modern snowmobiles take longer to cover the same distance. However, a British explorer named Tom Avery recently made the same trek in less than 37 days. In fact, Avery used the same kind of dogsled and the same number and breed of dogs as Peary had. Thus, Peary's claims are not impossible, and he very well might have been telling the truth.

Third, there are photographs taken by Peary that support his claim to have reached the North Pole. Measuring the shadows in Peary's photographs makes it possible to calculate the Sun's position in the sky. The Sun's position established from the photographs corresponds exactly to the Sun's position as it should have been at the North Pole on that day. This provides strong evidence that Peary reached the North Pole and took the photographs there.

Listening

There's no solid evidence that Robert Peary reached the North Pole. The arguments cited in the reading selection are not convincing.

First, it is true that the National Geographic Society committee declared that Peary had indeed reached the North Pole, but the committee was not completely objective. In fact, the committee was composed of Peary's close friends who had contributed large sums of money to fund Peary's trip. Moreover, the investigation lasted only two days. And according to Peary himself, the committee did not examine his records carefully. So the committee's conclusions seem biased and therefore are not trustworthy.

Second, the speed issue. Tom Avery's journey was different from Peary's in important ways. For example, Avery's sled was similar to Peary's sled, but Avery carried much less weight than Peary did, because Avery did not transport his food on the sled. Avery's food was dropped along the way by airplane. Moreover, Avery encountered highly favorable weather conditions, unlike Peary who travelled in very unfavorable conditions. So Avery's speedy trip was too different from Peary's to provide support for Peary's claims.

Third, the photographs do not prove anything. The techniques scientists use to determine the sun's position depend on measuring the shadows in the photographs very precisely. Without a precise measurement of the shadows, we cannot establish the sun's exact position. Now, Peary's pictures were photographed a hundred years ago using a primitive camera that took fuzzy, slightly unfocused photographs. Moreover, the photos have become faded and worn over time. As a result, the shadows in Peary's photographs look blurred and faded.

Those shadows cannot be used to calculate the position of the Sun with great accuracy. So we cannot be confident the photos were really taken at the North Pole.”

附录 2. TPO30 独立写作题目

Do you agree or disagree with the following statement?

It is more enjoyable to have a job where you work only three days a week for long hours than to have a job where you work five days a week for shorter hours.

Use specific reasons and examples to support your answer.

附录 3. TPO30 作文编号 07 和 18 的词汇替换对照表

No.	Sample 07	Sample 18
1	controversy	debate
2	erupts	arises
3	issue	topic
4	accurate	exact
5	applicant	seeker
6	opinion	view
7	optimal	best
8	employers	boss
9	comprehensive	good
10	websites	sites
11	provides	offers
12	additional	extra
13	information	info
14	unnoticed	ignored
15	credibility	trust
16	previous	former
17	several	some
18	drawbacks	problems
19	situation	case
20	allowing	letting