

2021

## 教育高危群体在写作过程上的性别差异研究

Randy E. Bennett  
*ETS*

Mo Zhang  
*ETS*

Sandip Sinharay  
*ETS*

Follow this and additional works at: <https://www.ce-jeme.org/journal>



Part of the [Educational Assessment, Evaluation, and Research Commons](#)

---

### Recommended Citation

Bennett, Randy E.; Zhang, Mo; and Sinharay, Sandip (2021) "教育高危群体在写作过程上的性别差异研究," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 2 : Iss. 1 , Article 2.

Available at: <https://www.ce-jeme.org/journal/vol2/iss1/2>

This Article is brought to you for free and open access by Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊. It has been accepted for inclusion in Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊 by an authorized editor of Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊.

# 教育高危群体在写作过程上的性别差异研究

Randy E. Bennett<sup>a</sup>, Mo Zhang<sup>a</sup>, and Sandip Sinharay<sup>a</sup>

<sup>a</sup> Educational Testing Service

## 摘要

本研究基于一个中学同等学力测验，考察了教育高危群体在写作过程上的性别差异。研究涉及了来自美国23个州的合计三万多考生，每一考生均参与了该语言测验的12副本考卷中的一个。研究借助键盘记录中抽取出的特征推断背后的写作过程，并将之整合为7个过程指标。研究结果发现女性被试的作文得分和语言测验总分均领先于男性，但领先程度很微弱。更重要的是，当控制了语言测验总分、年龄和作文题目后，全部7个过程指标均显示出显著的性别差异，其中，最突出的指标是流畅性和编辑性的不同方面。当前研究结果在许多重要方面与先前一些对在校生成人的研究结果相一致，也与在线和纸笔写作任务的研究结果相吻合。关于对使用字符类语言（如汉语）进行写作的个体如何开展类似研究，文章结尾给出了一些建议。

## 关键词

写作；  
问题解决过程；  
表现性评估；  
性别差异

已有研究明确表明，女性在写作测验上的表现优于男性，如美国国家教育进展评估 (the US National Assessment of Educational Progress, NAEP) 的结果所显示的。在NAEP的1988年到2011年的测试周期中，8年级和12年级的写作测试上均可观察到这种性别差异，且效果量达到中等程度 (Reilly, Neumann, & Andrews, 2019)。这种差异是很重要的，因为写作能力和水平与个体在美国高等教育系统中的成就直接相关 (Bridgeman & Lewis, 1994; Norris, Oppler, Kuang, Day, & Adams, 2006)，并且常常是就业市场上资方的要求。

随着写作测试的形式由纸笔形式发展到计算机化，我们能观察到的东西，也从结果拓展到了写作的过程。这些写作过程可以通过测试参加者的敲击键盘、停顿以及计算机测试中捕捉的其他事件和动作中推断出来。此外，对上述写作过程的研究，有助于研究者探索：各种人口学群体在写作过程中是否存在重要的差异。

当前研究主要考察了一个特定群体在写作过程上的性别差异：教育高危成人个体。这些个体的不利处境主要由受教育不足所致，因其未能完成中学教育，继而容易导致其低收入、低就业，并且伴随着更高比例的贫困和入狱可能 (Breslow, 2012; Scott, Zhang, & Koball, 2015)。

## 1 文献综述

写作能力的评判通常借助包含一个或多个写作题目的考试来进行，有时也辅以选择题。NEAP的写作

评估中，用上述方法评估群体的写作水平；托福考试 (TOEFL iBT) 和中国高考则用之评估个体写作水平。这些测试经常受到诟病，因为所测的限时的、按需的写作技能，明显有别于课堂论文写作中涉及的更深入的活动。即便如此，写作评估不仅能够预测个体在大学中的成就 (Bridgeman & Lewis, 1994; Norris et al., 2006)，并且其任务要求也并非截然不同于职场中的要求；在工作中，上级或客户通常要求个体迅速完成片段性的写作任务。事实上，写作技能是美国大学官员们最常用来表示学习成果的指标之一 (Hart Research Associates, 2016)。

测验情境下的研究表明，写作过程关系到作文的质量 (Bennett, Zhang, Deane, & van Rijn, 2020; Deane & Zhang, 2015)。因此，我们收集了与测验情境下写作过程有关的文献，以便于理解写作结果上的个体差异和群体差异。

在测评情境下，许多研究考察了中学英语测验中得到的作文 (写作产品)。这些研究发现了各种与作文质量有关的过程性指标 (measures)。例如，过程流畅性特征 (fluency process features) 与作文质量存在正相关，如总体打字速率、平均连击长度 (average burst length, 停顿间的连续击打字母数)；而表明拼写或录入时存在困难的单词内停顿 (within-word pauses)、键间停顿 (inter-key interval) 则与作文质量存在负相关 (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Bennett et al., 2020; Deane & Zhang, 2015; Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Zhang, Hao, Li, & Deane, 2018)。

其他能够预测作文质量的指标还包括编辑程度 (the extent of editing; Tate & Warschauer, 2019), 写作花费的总时长和写作过程中敲打的字数 (Bennett et al., 2020; Deane & Zhang, 2015), 以及写作计划程度 (planning), 比如语句之间的停顿程度 (Bennett et al., 2020; Deane & Zhang, 2015)。

现有研究不仅表明过程指标与作文质量有关, 还发现在控制作文质量后, 过程指标在不同人群间存在差异。有研究要求考生根据所给材料进行论证性写作, 结果发现低社经地位和非洲裔 (African American) 考生在效率上要分别低于相应的对照组 (高社经地位组和白人组[White]), 在同样数量的击键条件下写出的作文更短 (Guo, Zhang, Deane, & Bennett, 2019)。然而, 与当前研究关系更密切的是, 女性的流畅性要高于具有相似作文分数的男性。女性打字更快、使用更为复杂的单词、花费更长的时间写作、更快更频繁地编辑和停顿。另一研究同样要求借助给定材料进行写作并控制作文分数, 结果再次发现女性更流畅、编辑更深入, 在与计划有关的位置上停顿更少 (Zhang, Bennett, Deane, & van Rijn, 2019)。

写作认知理论和研究为上述结果给出了解释。上述工作任务表明工作记忆受限条件下, 低水平加工和高水平加工间存在竞争 (Kellogg, 2001; McCutchen, 1996, 2011)。当低水平加工过程, 如将想法转化为单词、句子、打字和拼写可以自动化进行时, 写作者能把注意力投注到更高水平的计划和修订上, 从而生成较高质量的文本。

研究表明, 相较于初学者, 熟练的写作者确实能够借助更长的组块流畅地写作; 他们花费更长时间计划、产出文本和修改, 倾向于在自然计划节点上停顿, 如从句和句子间 (Connelly, Dockrell, Walter, & Critten, 2012; Kaufer, Hayes, & Flower, 1986; McCutchen, 2006; McCutchen, Covill, Hoyne, & Mildes, 1994)。一些认知研究通过研究键盘记录再次确认了上述发现 (Baaijen, Galbraith, & De Gloppe, 2012; Leijten & Van Waes, 2013; Strömquist, Holmqvist, Johansson, Karlsson, & Wengelin, 2006; Sullivan & Lindgren, 2006; Wengelin, 2006)。

总之, 写作测评所测的技能与高等教育中的学业成就以及工作中的成功密切相关。在初中和高中中的写作考试中, 美国女生的水平高于男生。对具有相似写作分数的中学女生和男生的研究表明, 在基于给定材料进行说服性写作时, 这些群体使用了明显不同的方式。

当前研究旨在探索教育高危成人群体内是否同样存在上述认知过程差异, 主要研究问题有二:

1. 在教育高危成人群体内, 是否存在写作过程上的性别差异? 具体而言, 写作过程的流畅性、编辑

性上是否存在性别差异, 或认知理论和现有研究表明的其他行为差异?

2. 如果存在性别差异, 这些差异是否单纯为写作技能水平上的不同所致? 也就是说, 在控制了其他因素, 包括写作质量后, 群体相关过程 (group-related process) 的差异是否仍然存在?

## 2 方法

### 2.1 被试

本研究的被试来自一个测验项目, 该项目包括纸笔和计算机两种形式, 目的是核定未完成中学教育的个体是否具备中学同等学力, 通过测验者可获得相应证书。本研究选取了2017年9月到2018年8月的所有参加计算机版语言 (language arts) 测验写作分测验的考生。在该测验周期中, 写作分测验施测了12个副本考卷, 美国23个州合计得到32164条测验记录。我们移除了一些记录, 包括重复参加者的非首次记录、键盘记录损坏的记录, 过程评价无意义的记录 (空白作答或跑题导致的零分卷), 最终得到了30788名考生记录, 均为唯一作答。

出于分析目的, 我们将样本随机分为主样本和校验样本 (replication sample)。主样本和校验样本的人口统计特征见表1。从表1可知, 男性被试比例略高于女性 (53% 对 47%)。被试平均年龄为 24.5, 范围从十几岁的青少年一直到老年人。大约半数被试为白种人 (White), 其次为非洲裔 (African American, 16%) 和拉丁裔 (Hispanic, 主样本和校验样本分别为15%和16%)。90%以上的被试报告自己最好的交际语言为英语。大约三分之一的被试的最高教育水平为10年级或者低于10年级, 11年级和12年级的被试比例接近, 还有三分之一的被试未回答该题。关于母亲的最高教育水平, 有36%的被试未予作答, 接近五分之一的回答说具有高中学历, 还有五分之一回答说母亲具有大学肄业或大学学历。最后, 大约三分之一的被试报告自己从事全职或兼职工作, 超过三分之一的被试处于失业状态或不属于劳动力群体, 30%的被试未作答该题。

表2给出了不同性别群体的人口统计特征。从表2可知, 男女两个样本总的看来是相似的, 最大差异是男性被试在某些变量上缺失比例更高 (如种族、最高教育水平、母亲最高教育水平、工作状态等)。

这些缺失数据提示我们, 要慎重看待观察到的被试在同一变量上的微小差异 (如女性中有更高比例的白人, 但男性在此问题上有更高比例的缺失值)。年龄这个变量则是一个例外, 因为当前研究要求提供出生日期; 结果显示女性被试比男性平均年长两岁。

表 1  
样本的人口统计学特征

背景变量		主样本		校验样本	
		(N = 15,368)		(N = 15,420)	
年龄		24.5	(SD=9.1)	24.5	(SD=9.1)
性别					
	女	7193	(47%)	7231	(47%)
	男	8175	(53%)	8189	(53%)
种族					
	白种人 (White)	7527	(50%)	7421	(48%)
	亚裔 (Asian)	221	(1%)	199	(1%)
	非洲裔 (Black or African American)	2425	(16%)	2412	(16%)
	美国印第安人或阿拉斯加人 (American Indian or Native Alaskan)	258	(2%)	254	(2%)
	夏威夷或太平洋岛国人 (Native Hawaiian or Pacific Islander)	40	(0.3%)	32	(0.2%)
	拉丁裔 (Hispanic)	2336	(15%)	2419	(16%)
	多元种族 (Multiracial)	644	(4%)	633	(4%)
	其他种族 (Other race)	231	(2%)	230	(2%)
	未作答	1686	(11%)	1820	(12%)
英语是否最佳交际语言					
	是	14,310	(93%)	14,343	(93%)
	否	974	(6%)	991	(6%)
	缺失	84	(0.6%)	86	(0.6%)
最高教育水平					
	低于9年级	786	(5%)	764	(5%)
	9年级	1500	(10%)	1567	(10%)
	10年级	2878	(19%)	2868	(19%)
	11年级	4013	(26%)	3998	(26%)
	12年级	1116	(7%)	1098	(7%)
	缺失	5075	(33%)	5125	(33%)
母亲最高教育水平					
	小学	495	(3%)	526	(3%)
	高中肄业	1455	(9%)	1477	(10%)
	高中毕业	2858	(19%)	2831	(18%)
	大学肄业	1705	(11%)	1594	(10%)
	大学专科	524	(3%)	566	(4%)
	大学本科	735	(5%)	750	(5%)
	研究生	478	(3%)	446	(3%)
	未知	1563	(10%)	1570	(10%)
	缺失	5555	(36%)	5660	(37%)
工作状态					
	兼职	2248	(15%)	2298	(15%)
	全职	2757	(18%)	2736	(18%)
	失业	3750	(24%)	3682	(24%)
	非劳动力	2030	(13%)	2014	(13%)
	缺失	4583	(30%)	4690	(30%)

表 2  
不同性别的人口统计学特征

背景变量	男 ( <i>N</i> = 16,364)	女 ( <i>N</i> = 14,424)
年龄	23.5 ( <i>SD</i> =8.7)	25.6 ( <i>SD</i> =9.4)
种族		
白种人 (White)	7856 (48%)	7092 (49%)
亚裔 (Asian)	185 (1%)	235 (2%)
非洲裔 (Black or African American)	2505 (15%)	2332 (16%)
美国印第安人或阿拉斯加人 (American Indian or Native Alaskan)	270 (2%)	242 (2%)
夏威夷或太平洋岛国人 (Native Hawaiian or Pacific Islander)	33 (0.2%)	39 (0.3%)
拉丁裔 (Hispanic)	2358 (14%)	2397 (17%)
多元种族 (Multiracial)	699 (4%)	578 (4%)
其他种族 (Other race)	256 (2%)	205 (1%)
未作答	2202 (13%)	1304 (9%)
英语是否最佳交际语言		
是	15,258 (93%)	13,395 (93%)
否	1001 (6%)	964 (7%)
缺失	105 (0.6%)	65 (0.5%)
最高教育水平		
低于9年级	694 (4%)	856 (6%)
9年级	1403 (9%)	1664 (12%)
10年级	2757 (17%)	2989 (21%)
11年级	4162 (25%)	3849 (27%)
12年级	1167 (7%)	1047 (7%)
缺失	6181 (38%)	4019 (28%)
母亲最高教育水平		
小学	357 (2%)	664 (5%)
高中肄业	1193 (7%)	1739 (12%)
高中毕业	2814 (17%)	2857 (20%)
大学肄业	1606 (10%)	1693 (12%)
大学专科	543 (3%)	547 (4%)
大学本科	830 (5%)	655 (5%)
研究生	496 (3%)	428 (3%)
未知	1722 (11%)	1411 (10%)
缺失	6803 (42%)	4412 (31%)
工作状态		
兼职	2099 (13%)	2447 (17%)
全职	2860 (17%)	2633 (18%)
失业	3863 (24%)	3569 (25%)
非劳动力	1882 (12%)	2162 (15%)
缺失	5660 (35%)	3613 (25%)

## 2.2 工具

当前研究中涉及的中学同等学力测验的每一副本考卷均包含五个分测验。考生可以在一天内完成所有分测验,也可以按照自己喜好以任意顺序在不同天完成测验。阅读分测验包含40个多项选择题,用以考察被试理解、领会、解释和分析多种阅读材料的能力。写作分测验包含50个多项选择题和一个作文题,用以测量被试在写作过程中编辑和修改文本,生成和组织观点的能力。第三个分测验是数学分测验,包含50个多项选择题,主要考察基本概念和推理技能。科学分测验也包含50个多项选择题,主要考察个体的科学知识水平、科学探究原理的运用,以及对科学信息的解释和评价。最后一个分测验是社会学分测验,同样包含50个多项选择题,主要用于考察被试在分析和评价领域相关信息方面的技能。

每个分测验的量尺分取值范围均为1—20。作文题作为写作测验的一部分,还报告原始分,取值范围为0—6。大多数州或者领土地区设定的通过标准均要求:任一分测验的最低得分不低于8分,作文得分不低于2分,测验总分(五个分测验合计)不低于45分。

作文题是本研究的关注点,该题要求考生阅读关于同一主题的阐述视角不同的两篇文章(例如,成功更多地取决于天赋还是努力)。这两篇文章通常长约750—800词,并且长度基本相等。考生需要写出一篇作文,文中要运用上述文章中得到的证据来阐释自己的观点。

## 2.3 评分

作文评分由职业评分员完成,这一过程属于该测验项目的操作性评分过程的环节之一。评分之前用一个作文真题对评分员进行训练,根据评分员对该题的打分判断其是否具备资格;进入正式评分后每天都基于操作性样题进行校准。评分过程中,组长会对评分员的操作性评分表现进行监控,主要借助评分序列中加入的效度样本,当评分员对效度样本的给分偏离了专家的预先给分时,组长就会对其加以干预。

评分细则包含四方面内容(中心观点的展开、观点的组织、语言功用和写作规范),评分员要据此进行综合评分。对于给定的作文题目,评分细则会附有该题相关的注释,并标注出每一分值的反应范围。

每一考生作答均由两名评分员评分。当两名评分员的给分差异大于1分时,该作答交由第三名评分员进行评分。

当前数据集在去除了零分试卷后,两名评分员在12个作文题上评分一致比例的中位数为78%(范围从76%到79%),平方加权kappa系数的中位数为0.65(范围从0.61到0.68)。本研究主要采用了前两名评分员给

分的均值。

我们从每个考生作答中提取了78个过程特征,这些过程特征变量选自一个多达数百的特征集,这一特征集来自一个关于写作过程分析的大型研究项目(Deane, 2014; Deane & Zhang, 2015)。这78个特征包括计数(counts,即总击键次数),持续时间(durations,即连续按击退格键的最小时间间隔),速率(rates,即远距离编辑跳转动作的最大速率)和拼错概率(probabilities,最终文本中的某单词曾被拼错的相对概率)。接下来对这些特征进行探索性因子分析,以便将其编制成过程量表。对12个题目分别进行基于主成分分析的探索性因子分析,并进行斜交旋转。如果某个特征在12个作文题的至少8个上满足同一因子的同一方向上的载荷高于.30,就予以保留。这一原则使得得到的特征集在题目间具有合理的一致性。为了兼顾结果的解释性和简洁性,研究最终抽取出7个因子,合计解释了62%的总变异。该7因子涉及了59个特征,其中51个(86%)满足11或12题的选择标准,5个满足10题,2个满足9题,1个满足8题。

基于上述过程我们编制了7个过程量表(process scale),每一量表包含那些在相应因子上具有一致性载荷的特征。对于每一量表,我们将每一被试在每一特征上的得分标准化(均值为0,标准差为1)后求和,得到一个过程指标(process indicator)。因为每个过程指标由不同数目的特征构成,所以无法进行直接比较。

上述过程指标及其相应特征,以及得自因子分析的每一特征的效价(valence)详见附录。过程指标的命名主要根据其优势特征。指标一为单词间速度(Between-Word Speed),包含九个特征,如单词间敲击空格键速率(the rate of spacing keystrokes between words,正效价)和间隔持续时间(pause durations,负效价)。这一指标可以视作过渡到下一单词的功能指征。较快的过渡表明在词汇提取(即单词查找)、句法编码(即按恰当顺序安排单词)和录入功能的某种较大指令的联合。第二个指标是大连击流畅性(Large-Burst Fluency),包含11个特征。这一指标合并了多种特征,如长片段文本连击长度(burst length for long segments of text,正效价),以及录入精确性(typing precision),即最终文本中的某个单词的拼错概率(负效价)。这些特征合并在一起,表示相对较长的不含错误的连击的自动性,不管是否经过纠错。指标三称为产出性(Productivity),涉及8个特征,反映了个体的文本产出量(尽管这些文本不一定留在最终提交的作文中)。该指标包含的特征有区间内击键总数(the total number of keystrokes in the session,正效价)、花费时间去重新安排文本的概率(the probability of time being spent rearranging text,负效价,较高的概率性意味着较低的效率)。故此,产

出性表征着流畅性和效率。指标四删除性编辑 (Deletion Editing) 包含9个特征, 所涉变量主要与文本的剪切和删除有关(如用时 *time spent*、剪切掉的字母数 *number of characters cut*、远距离跳跃的最大速率 *maximum rate of long-distance jumps* 等), 所有变量均具有正效价。这一指标可用于表征个体进行多次快速剪切的倾向, 包含那些远离当前鼠标位置的剪切。指标五为跳跃性编辑 (Jump Editing), 指的是远距离的文本改变, 需要从当前鼠标位置到一个非相邻位置, 可能是同一单词或同一个句子内部, 也可能是上一句或者再之前的句子。这一指标包含12个特征, 如跳跃性编辑用时 (*the time spent on jump edits*), 击键的最长持续时间 (*the maximum duration in keystrokes*), 这些特征均具有正效价。跳跃性编辑表明了单词、短语、句子层面的文本监控, 有时也可能是文本正面的监控。指标六称为句间过渡和退格速度 (*Between-Sentence Transition and Backspace Speed*), 包含六个特征, 如句间空格击键的最大速率 (*the maximum rate of between-sentence space keystrokes*, 正效价) 和连续退格间持续时间的最小值 (*the minimum duration of continued backspaces*, 负效价)。这一与流畅性相关的指标看似主要表示从一个句子到另一个句子的流畅程度, 以及借助退格键快速移除已经写好的文本的程度。最后四个段落特征 (*Paragraphing features*) 均有正效价, 涉及与行间停顿有关的击键数目和持续时间。这一指标可能表示在段落间进行文本组织和计划的努力程度。

## 2.4 数据分析

我们首先将被试样本随机分成两部分, 样本一为主样本, 样本二为校验样本。在主样本进行的所有分析在校验样本均重复进行一次, 某一结果在两个样本上均达到显著才具有意义。

为判定研究群体在写作过程上是否存在差异, 以性别为自变量, 每一过程指标上的分数为因变量, 进行多元方差分析 (MANOVA)。首先进行了单变量检验 (*univariate tests*), 以检测因变量在不同性别间是否差异显著。接下来进行多元方差分析, 自变量和因变量不变, 但将写作总分、作文题目 (哑变量) 和年龄作为协变量 (同时加入二阶交互作用项)。这一分析意在判定所观察到的过程差异是否仅源自不同性别群体在这些变量 (即写作技能、年龄或者作文题目) 上的差异。

所有分析均借助SAS程序完成, 其中因子分析使用 *proc factor* 模块, MANOVA/MANCOVA使用了 *proc glm* 模块 (Pett, Lackey, & Sullivan, 2003; Kuehl, 2000)。所有分析均使用双侧检验, 大多数情况下  $\alpha$  水平设置为 .05。对于MANOVA/MANCOVA后的单变量检验, 使用了Bonferroni校正, 置信水平设为  $\alpha/7$  (7为过程指标数

= .0071。此处使用校正正是因为上述检验与研究问题直接相关。

## 3 结果

### 3.1 总结性统计

表3给出了高中同等学力测验各个分测验的分数均值。不同性别组在语言测验的阅读分测验上并无显著差异, 但在写作分测验上女性被试的得分显著高于男性。然而, 按照 (Cohen, 1988) 的分类标准 (即  $<0.20$  标准差单位), 上述性别差异是很微弱的。与之形成对照的是, 男性被试在其余三个分测验上的得分均显著高于女性, 包括数学、科学和社会学, 尽管所有差异都相对较小 (在0.22到0.36个标准差单位间)。

表3  
不同性别被试中学平等学历测验得分均值

	主样本		校验样本	
	男	女	男	女
	写作 (量尺=1-20)			
平均数	11.40	11.80**	11.30	11.60**
标准差	3.39	3.27	3.36	3.30
	阅读 (量尺=1-20)			
平均数	11.80	11.80	11.70	11.70
标准差	4.18	4.12	4.22	4.15
	数学 (量尺=1-20)			
平均数	11.90	10.60**	11.80	10.50**
标准差	4.54	4.63	4.59	4.62
	科学 (量尺=1-20)			
平均数	13.90	13.00**	13.80	12.90**
标准差	4.00	3.90	4.04	3.87
	社会学 (量尺=1-20)			
平均数	13.70	12.20**	13.70	12.10**
标准差	4.49	4.50	4.44	4.51

注: 男女群体样本量从6,253到8,289不等。

\*\* $p < .0001$ .

表4给出了作文平均得分、长度和作答时长。在这三项指标上, 不同性别群体间的差异是显著的。女性的作文得分高于男性, 作文也长于男性, 花费的时间也更久一些。当然, 所有差异都是很微弱的<sup>1</sup>。

表5给出了过程指标间的相关系数和内部一致性信度系数 (这些系数或者直接给出或者可根据已给数据计算得到)。表5同时给出了主样本和校验样本的分析

<sup>1</sup>类似的微弱差异在的语言测验阅读分测验的多项选择题部分也被发现了, 这意味着性别差异的大小与题目类型无关。

表4  
不同性别被试作文得分、长度和用时情况

	主样本		校验样本	
	男	女	男	女
人数	8175	7193	8189	7231
作文得分(量尺=1-6)				
平均数	2.60	2.80**	2.60	2.70**
标准差	0.89	0.86	0.89	0.87
作文长度(单词数)				
平均数	288.00	308.00**	289.00	301.00**
标准差	130.00	131.00	131.00	129.00
时长(分钟)				
平均数	33.70	34.50*	33.50	34.30*
标准差	16.40	15.90	16.20	16.10

\* $p < .005$ . \*\* $p < .0001$ .

结果,所有结果在统计上都是显著的,且两个样本间仅有微弱差异。从表5可知,过程指标具有较高的信度( $\alpha$ 系数从.87到.96)。此外,过程指标间的相关较低(主样本从.01到.46,校验样本从.01到.45),说明这些指标在一定程度上彼此相对独立。

在所有过程指标中,产出性特征与其他指标的相关最高。这一指标可以视为表示流畅性的指标,涉及的特征包括单词间速度( $r = .42$ )、句间过渡和退格速度( $r = .41$ )等,因为在这些过渡上用时越少,留给用于产出文本的时间就越多。产出性还与编辑性具有中等大小的相关,如与跳跃性编辑的相关为.40以上,与删除性编辑的相关为.37,这很可能是由于高产产出性使得有更多的文本用于编辑。这一指标与作文长度之间的高相关( $r = .83$ )可以说明上述推测的合理性。最后,产出性与用时相关在.50以上,与作文得分相关达到.68,与语言测验写作总分相关达到.42。

此外,表5结果中还应注意的是,所有过程指标与作文得分和语言测验写作总分均具有正相关。包括产出性在内,各项指标与作文得分的相关分布在.17到.68间,与写作总分的相关则分布在.19到.42间。与作文得分和写作总分相关第二高的指标是单词间速度,相关分别为.38和.37。跳跃性编辑也和上述两个得分具有明显相关,分别为.34和.30。大连击流畅性与作文得分的相关最低( $r = .17$ ),与写作得分相关次低(主样本为.21,校验样本为.20)。这一特征与其他指标间的相关也很低(而且总是负值),并且与用时的相关也较小,说明能准确地写出较长文本的考生略倾向于使用更少的时间。

最后,正如我们所预期的那样,过程指标与语言

测验写作总分的相关要高于与数学测验分数的相关。然而,尽管这些指标与数学测验分数的相关较低,但其水平近似于社会学、科学以及阅读测验分数的相关。这一结果也许意味着,对于所有测量指标,有一个共同的流畅性因子在起作用。事实上,分析结果中相关最高的是单词间速度和产出性与数学测验分数间的相关,而这两个指标都与流畅性有关。

### 3.2 研究问题1

为了考察不同性别群体在写作过程上是否存在差异,我们比较了不同群体在每一过程指标上的平均得分(见表6)。分析结果发现,写作过程上存在显著的性别差异,主样本的MANOVA检验统计量 Pillai's Trace = 0.045,  $F(7, 15360) = 103.77, p < .0001$ ; 校验样本的 Pillai's Trace = 0.043,  $F(7, 15412) = 98.09, p < .0001$ 。单因素方差分析结果显示,除了大连击流畅性外( $p = .2710$ ),不同性别群体在其余所有指标上均差异显著( $p < .0001$ )。具体说来,女性群体在单词间的过渡更迅速,产出更长的文本,进行更多的删除性编辑和跳跃性编辑,更迅速地进行句间过渡和退格,更努力地进行组织,甚至段落间计划等。

### 3.3 研究问题2

因为写作过程上的组间差异可能是由其他变量差异所致,进一步分析中将语言测验写作总分、年龄和作文题作为协变量,再次比较男女两组在过程指标上的差异。分析结果再次表明男女群体差异显著,其中主样本的 Pillai's Trace = 0.055,  $F(7, 15347) = 128.36, p < .0001$ , 校验样本的 Pillai's Trace = 0.052,  $F(7, 15399) = 121.79, p < .0001$ <sup>2,3</sup>。与MANOVA分析结果不同的是,此处的单变量检验在所有七个过程指标上都性别差异显著,包括大连击流畅性,其结果也变得显著了。

指标的调整均值和效果量见表7。值得注意的是,协变量的调整效应通常最小。更重要的是,进一步检验效果量的结果表明,有四个过程指标上的性别差异值得注意:单词间速度(主样本  $d = 0.39$ , 校验样本  $d = 0.35$ ),产出性(主样本  $d = 0.30$ , 校验样本  $d = 0.33$ ),删除性编辑( $d = 0.30$  和  $0.33$ ),以及跳跃性编辑( $d = 0.28$  和  $0.26$ )。按照Cohen(1988)的分类标准(即0.20到0.49),这些效应属于较小的效应。剩下的指标只有0.14或者更小的效应,只能称为微乎其微了。

<sup>2</sup>语言测验分数和年龄在所有的指标上均存在显著的主效应,作文题在单词间速度、大连击流畅性、产出性和删除性编辑上主效应显著。

<sup>3</sup>当在模型中加入所有二阶交互作用项时,并非所有交互作用项均显著,因此在此只报告了主效应。



表5  
研究变量间的相关

变量	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. 单词间速度	<b>.96</b>	-.03	.42	.25	.07	.34	.10	.38	.37	-.15	.35	.28	.24	.22	.27
2. 大连击流畅性	-.01	<b>.94</b>	-.03	-.12	-.15	-.10	-.02	.17	.20	-.22	.14	.18	.10	.15	.15
3. 产出性	.42	-.01	<b>.94</b>	.37	.45	.41	.30	.68	.42	.55	.83	.30	.26	.24	.27
4. 删除性编辑	.24	-.12	.37	<b>.93</b>	.32	.21	.17	.25	.25	.31	.22	.20	.17	.15	.17
5. 跳跃性编辑	.08	-.14	.46	.33	<b>.87</b>	.20	.28	.33	.19	.59	.36	.14	.09	.10	.11
句间过渡和退格速度															
6. Between-sentence transition & Backspace Speed	.33	-.10	.41	.20	.21	<b>.87</b>	.17	.24	.22	.22	.21	.18	.15	.14	.15
7. 段落特征	.11	-.03	.28	.17	.30	.18	<b>.94</b>	.30	.21	.27	.26	.17	.14	.15	.16
8. 作文得分	.38	.17	.68	.25	.34	.24	.30	-	.60	.31	.74	.42	.35	.36	.38
9. 写作得分	.37	.21	.42	.24	.20	.20	.20	.60	<b>.71</b>	.07	.40	.71	.54	.65	.67
10. 作文用时	-.16	-.22	.54	.31	.58	.22	.26	.30	.06	-	.42	.08	.05	.07	.05
11. 作文长度	.34	.16	.83	.22	.37	.21	.24	.73	.40	.41	-	.27	.26	.23	.26
12. 阅读	.28	.16	.31	.18	.16	.17	.15	.42	.70	.09	.27	<b>.85</b>	.50	.70	.69
13. 数学	.25	.10	.25	.17	.08	.13	.13	.34	.53	.04	.24	.50	<b>.75</b>	.53	.61
14. 社会学	.22	.15	.24	.16	.11	.13	.14	.36	.65	.07	.22	.69	.53	<b>.83</b>	.70
15. 科学	.26	.15	.27	.18	.14	.15	.16	.38	.67	.07	.25	.69	.61	.70	<b>.85</b>

注：对角线上的黑体数字为基于当前数据计算得到的过程指标和测验手册中提供的同等学力测验分数（估计自2015年测试）的 $\alpha$ 值。对角线之下为主样本估计值，之上为校验样本估计值。主样本估计值绝对值高于.07的，校验样本估计值绝对值高于.04的在统计上显著地不为零 ( $p < .0001$ )。主样本人数从13,557 到15,368; 校验样本人数从13,618到15,420。

表 6  
不同性别被试的过程指标统计信息

性别	主样本			校验样本		
	平均数	标注差	取值范围	平均数	标注差	取值范围
单词间速度						
女	1.08	7.33	-25.12 to 23.52	0.89	7.39	-24.76 to 22.81
男	-0.95*	8.13	-25.13 to 23.52	-0.79*	8.10	-24.76 to 23.52
大连击速率						
女	-0.07	8.69	-19.88 to 36.41	-0.08	8.72	-19.89 to 36.33
男	0.06	8.92	-19.49 to 36.26	0.07	8.80	-19.35 to 35.12
产出性						
女	0.98	6.77	-13.40 to 26.04	0.78	6.75	-13.29 to 26.43
男	-0.86*	6.54	-13.40 to 26.04	-0.68*	6.58	-13.29 to 26.43
删除性编辑						
女	1.20	7.56	-6.64 to 30.69	1.28	7.65	-6.61 to 30.38
男	-1.06*	6.75	-6.64 to 27.07	-1.13*	6.60	-6.61 to 27.36
跳跃性编辑						
女	1.24	7.37	-29.24 to 26.99	1.14	7.56	-29.77 to 27.15
男	-1.09*	7.74	-29.44 to 24.10	-1.00*	7.56	-29.83 to 28.86
句间过渡和退格速度						
女	0.34	4.68	-9.52 to 15.78	0.17	4.65	-9.44 to 16.06
男	-0.30*	4.71	-9.52 to 15.78	-0.15*	4.73	-9.44 to 16.06
段落特征						
女	0.22	3.49	-9.80 to 5.40	0.17	3.57	-9.48 to 5.42
男	-0.19*	3.84	-9.80 to 5.40	-0.15*	3.81	-9.48 to 5.42

注：在进行男女差异的单因素F检验时，Bonferroni校正的显著性水平设为 $p < .0071$ 。其中，主样本包含男7,193人，女8,175人；校验样本包含男7,231人，女8,189人；对每一样本，所报告结果是12个作文题上的汇总版本。

## 4 讨论

本研究基于一个中学同等学力测验，考察了教育高危群体在写作过程上的性别差异。研究所用的样本涵盖了超过30,000个成年人，每人都参加了该语言测验的12个复本之一的测试，每个复本也都包含一道写作任务。研究中的被试分布遍及全美半数以上的州。本研究从键盘记录中抽取出特征，将这些特征整合为7个写作过程指标，并且使用了主样本和校验样本来确保研究结果具有意义。研究结果发现，女性被试在作文和语言测验上的得分均领先于男性，但领先程度微乎其微（即 $< 0.20$ 个标准差单位）。更为有趣的是，男女被试在7个过程指标中的6个上差异显著。当控制了作文总分、年龄和作文题目后，所有指标上的差异均变得显著了，其中效果量最大的指标为单词间速度、产出性、删除性编辑和跳跃性编辑。

教育高危的成年女性在作文和语言测验上的得分

仅稍稍高于男性，这一发现在某种程度上令人惊讶。从NAEP的结果可知，在多个测试周期中，八年级和十二年级在写作测试上均存在中等大小的性别差异(Reilly et al., 2019)。当前的中学同等学力测验除了测量写作的方式与NEAP的工具有所不同外，二者的主要差异在于所涉及的被试群体。NAEP测试的是年轻人的全国代表性样本，而当前研究样本处于能力分布的相对低端，平均年龄要比NEAP的大上几岁，并且考生群体内部年龄差异也要大得多。从NAEP结果我们还可以知道，在分布的低端，写作上的性别差异要大一些(Reilly et al., 2019)，因此提供了不同成分的男女群体选择参考同等学力测验的可能性。例如，男性入狱的概率要数倍于女性(Glaze & Kaeble, 2014, p. 6; IES, 2014)，而入狱者的平均读写水平通常要低于一般人群(IES, 2014)。即使许多监狱给犯人提供了参加同等学力测验的机会，我们还是有理理由认为犯人参加该测验的

表 7

主样本和校验样本中不同性别被试的过程指标调整均值和效果量（以年龄、语言测验作文总分和作文题目为控制变量）

性别	主样本		校验样本	
	平均数	效果量	平均数	效果量
单词间速度				
女	1.17		1.05	
男	-1.13*	0.39	-1.03*	0.35
大连击流畅性				
女	-0.34		-0.20	
男	0.25*	-0.08	0.33*	-0.08
产出性				
女	0.89		0.70	
男	-0.88*	0.27	-0.81*	0.23
删除性编辑				
女	1.16		1.17	
男	-1.01*	0.30	-1.21*	0.33
跳跃性编辑				
女	1.13		1.02	
男	-0.95*	0.28	-0.92*	0.26
句间过渡和退格速度				
女	0.32		0.17	
男	-0.33*	0.14	-0.23*	0.08
段落特征				
女	0.18		0.11	
男	-0.12*	0.08	-0.11*	0.06

注：在进行男女差异的单因素F检验时，Bonferroni校正的显著性水平设为 $*p < .0071$ 。其中，主样本包含男7,193人，女8,175人；校验样本包含男7,231人，女8,189人；对每一样本，所报告结果是12个作文题上的汇总版本。效果量等于女性平均分与男性平均分的差值除以总体标准差。

机会要低于一般人群，因此相当于从测试群体中去除了更多的低能力的男性被试。

影响研究结果的另一因素是同等学力测验的政策在不同州间有所不同。也就是说，有些州只提供唯一测验（如密苏里州），而有些州允许考生在两项或更多测验中选择一个参加（如加利福尼亚州）。此外，有些州允许考生选择测验的形式，即计算机形式或纸笔形式。在能够有所选择的时候，男女考生很可能选择参加的测验或测验形式有所区别。我们对比了提供唯一测验和提供多个测验的州的情况，结果发现前者在作文和语言测验作文总分上的性别差异要大一些<sup>4</sup>。此外，尽管过程指标上那些重点的性别差异的模式实际

上并未改变，但总的说来，单个过程指标的效应在仅提供单一工具的州要更大一些，意味着当前的全样本结果对教育高危成年人的性别差异可能有所低估<sup>5</sup>。这一点对单词间速度和产出性而言尤为如此，二者各自的效果量分别为0.17和0.19，但对于单一工具州则分别为0.31和0.31。

本研究所探查到的写作过程差异有何深层含义呢？在展现出明显性别差异的指标中，单词间速度和产出性均直接或间接地与流畅性有关，前者主要度量移向下一单词的快速性，后者则直接表示能否快速地产出和录入相当量的文本。这两个指标具有重要的意义，

<sup>4</sup>作文分数的效果量（以州N加权）平均数在多工具州（ $N = 13,650$ ）和单一工具州（ $N = 17,135$ ）分别为0.09和0.19。在作文和语言测验得分上，效果量的平均效果量为0.02和0.17。

<sup>5</sup>多工具州和单一工具州的过程指标分数的平均效果量分别为：单词间速度0.17和0.31，大连击速率0.01和0.01，产出性0.19和0.31，删除性编辑0.31和0.36，跳跃性编辑0.29和0.30，句间过渡和退格速度0.07和0.13，段落特征0.10和0.11。

因为它们既和作文分数相关(单词间速度 $r = .38$ ; 产出性 $r = .68$ ), 又和语言测验作文总分相关(单词间速度 $r = .37$ ; 产出性 $r = .42$ )。删除性编辑和跳跃性编辑则与监控和修改有关, 同时也与写作流畅性相关, 尽管相关较低( $r$ 从.19到.34)。

本研究所发现的性别差异不仅与写作质量有关, 而且类似于先前在一些参加某在线给材料写作测试的中学生身上发现的结果。Guo et al. (2019) 研究发现, 当控制了作文分数后, 女性被试比男性打字快, 花更长时间产生文本, 更迅速和频繁地编辑以及停顿。在另一个被试样本中, Zhang, Bennett, et al. (2019) 则发现女性被试比男性更流畅, 更喜欢编辑, 与计划相关的停顿更少, 继而表明她们具有更高的流畅性。最后, 较早的一些研究曾使用更简单的基于纸笔的写作任务, 结果发现从小学生到成年人, 在流畅性和产出性上都具有相当程度的性别差异(Camarata & Woodcock, 2006; Jewell & Malecki, 2005; Malecki & Jewell, 2003)。相应地, 当前研究结果也表明写作过程上的性别差异同样存在于计算机考试场景下参与同等学力测验的高危成年群体。

上述过程差异对学业成就和教学有何意义呢? 潜在的重要启示之一与写作过程的性质有关。正如本研究以及其他研究均发现的那样(Bennett et al., 2020), 更高水平的流畅性和产出性通常对应着更高水平的作文表现。理论认为, 更高水平的流畅性或基本过程的自动化加工, 如单词提取和录入, 使得能够将认知资源投入到更高阶的过程中, 如观点生成和句子计划(McCutchen et al., 1994; McCutchen, 1996)。此外, 编辑的程度关系到文本的质量(Tate & Warschauer, 2019; Zhang, Zhu, Deane, & Guo, 2019)。在控制了写作技能后, 在流畅性和编辑相关的指标上仍然可观察到明显的性别差异, 这意味着男性被试在这些过程中可对其劣势进行补偿。当前研究中, 男性被试在句子间过渡和退格速度上低于女性(边缘显著), 表明男性在这些边界位置上可能需花费更多时间去产出下一步要表达的内容。在对中学生的研究中, 相似的现象发生在文本生成爆发间的停顿时间上, 包括句子间停顿; 研究表明这些停顿通常反映了背后的计划行为(比如, Révész, Michel, & Lee, 2019)。女性由于在流畅性上具有优势, 因此在这些位置上的停顿需求较少, 在写作时能够更高效地进行计划。无论如何, 识别男性所用的补偿机制或许有助于低分考生提高其写作水平。其他一些方法则给男性被试回放其写作过程, 并请其描述自己当时所想所做, 这些方法或许有助于对男性所用补偿机制的识别。

研究表明, 无论在线或纸笔写作测试上, 女性均具有流畅性优势。这一结果意味着键盘录入能力本身

可能并不是导致差异的原因。更可能的解释是在生理、社会和心理因素的联合影响下(Reilly et al., 2019), 女性具有几分更高的言语天赋(Halpern, 2012)。为了比较言语流畅性和录入技能的相对贡献, 进一步的研究将同时施测写作任务和单纯的录入(拷贝)任务(Deane et al., 2018), 或者计算易提取的高频单词的输入速率(Zhang, Deane, Feng, & Guo, 2019)。

至于研究意义, 我们感兴趣的是, 本研究中发现的写作过程上的性别差异是否可以拓展到其他语言呢? 包括哪些基于字符的语言, 如汉语。对中国儿童的纸笔测试研究表明, 句子水平的手写流畅性和性别与写作整体质量相关(Yan et al., 2012)。然而, 网络环境中进行写作流畅性评价面临特殊的挑战, 因为汉字不象罗马字母那样直接和键盘上的键一一对应。在最常用的输入法中, 录入汉字主要借助在全键盘上录入拼音。因为汉语中包含的同音字很多, 许多汉字具有相同的拼音, 导致录入者必须在一组可能汉字中选出目标汉字(例如, 输入拼音“wu”会得到数个汉字, 含义各个不同, 包括五、午、务、无、舞、物、吴、乌和伍等)。此外, 在拼音录入结束前, 各个同音字的可能性就展现在录入者眼前, 甚至可能自行完成了选择。输入法面板对上下文和录入者的个人偏好都很敏感, 会按照与先前文本的拟合度排列可能性。另一种五笔输入法则基于汉字结构, 而不是其发音。五笔输入法中, 罗马字母与汉字的的不同部分相匹配, 每个汉字至多对应四个字母, 这种设计使得五笔打字通常快于拼音输入法。在测量写作过程时, 个体所用输入方法的差异必须考虑在内。例如, 可以拟合混合分布, 将流畅性作为录入方法的函数加入模型。

本研究主要存在以下局限。首先, 研究所选取的写作任务是基于测验的单一体裁的任务。这种任务可能明显不同于教育高危成年群体在继续教育或工作中面临的写作任务。其次, 选取的过程测量指标仅是有效写作必须指标的一个子集。这些指标尤其指向如流畅性等基础过程, 而不是如观念生成等高阶过程。最后, 研究对教学的直接启示相对有限。要拓展写作过程数据在教与学促进上的应用前景, 更多工作仍然有待开展。

## 参考文献

- Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (Research Report RR-12-23). Princeton, NJ: Educational Testing Service.
- Baaijen, V. M., Galbraith, D., & De Glopper, K. (2012). Keystroke analysis: Reflections on procedures and

- measures. *Written Communication*, 29(3), 246–277.
- Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment*, 1–20.
- Breslow, J. M. (2012, September). By the numbers: Dropping out of high school. *Frontline*.
- Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, 31, 37–50.
- Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, 34, 231–252.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Connelly, V., Dockrell, J. E., Walter, K., & Critten, S. (2012). Predicting the quality of composition and written language bursts from oral language, spelling, and handwriting skills in children with and without specific language impairment. *Written Communication*, 29, 278–302.
- Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report RR-14-03). Princeton, NJ: Educational Testing Service.
- Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018). *Behavioral differences between retyping, drafting, and editing: A writing process analysis* (Research Memorandum RM-18-06). Princeton, NJ: Educational Testing Service.
- Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (Research Report RR-15-26). Princeton, NJ: Educational Testing Service.
- for Education Sciences (IES), I. (2014). *US PIAAC prison study results: 2014* (Tech. Rep.). Retrieved from [https://nces.ed.gov/surveys/piaac/results/prison\\_summary.aspx](https://nces.ed.gov/surveys/piaac/results/prison_summary.aspx)
- Glaze, L. E., & Kaeble, D. (2014). *Correctional populations in the United States, 2013*. Washington, DC: US Department of Justice. Retrieved from [https://www.aacu.org/sites/default/files/files/LEAP/2015\\\_Survey\\\_Report3.pdf](https://www.aacu.org/sites/default/files/files/LEAP/2015\_Survey\_Report3.pdf)
- Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, 55, 194–216.
- Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, 44, 571–596. <https://doi.org/10.3102/1076998619856590>.
- Halpern, D. F. (2012). *Sex differences in cognitive abilities*. New York: Psychology Press.
- Hart Research Associates. (2016). *Trends in learning outcomes assessment. Key findings from a survey among administrators at AAC&U member institutions* (Tech. Rep.). Retrieved from [https://www.aacu.org/sites/default/files/files/LEAP/2015\\_Survey\\_Report3.pdf](https://www.aacu.org/sites/default/files/files/LEAP/2015_Survey_Report3.pdf)
- Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, 34(1), 27–44.
- Kaufert, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, 20, 121–140.
- Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, 114(2), 175.
- Kuehl, R. O. (2000). *Design of experiments: statistical principles of research design and analysis* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30, 358–392.
- Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools*, 40(4), 379–390.
- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, 8(3), 299–325.
- McCutchen, D. (2006). Cognitive factors in children's writing. In C. A. MacArthur, S. Graham, & F. Jill (Eds.), *Handbook of writing research*. New York: Guilford Press.
- McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research*, 3(1), 51–68. <http://dx.doi.org/10.17239/jowr-2011.03.01.3>.
- McCutchen, D., Covill, A., Hoyne, S. H., & Mildes,

- K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology, 86*(2), 256–266. <https://doi.org/10.1037/0022-0663.86.2.256>.
- Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT Writing validation study: An assessment of predictive and incremental validity* (Research Report 2006-2). New York: College Board. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.189.5460&rep=rep1&type=pdf>
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications Inc.
- Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445–458.
- Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition, 41*, 605–631.
- Scott, M. M., Zhang, S., & Koball, H. (2015). *Dropping out and clocking in: A portrait of teens who leave school early and work* (Low-Income Working Families Brief). Washington, DC: Urban Institute. Retrieved from <https://www.urban.org/sites/default/files/publication/49216/2000189-Dropping-Out-and-Clocking-In.pdf>
- Strömquist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, Å. (2006). What keystroke-logging can reveal about writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing* (pp. 45–71). Oxford, UK: Elsevier.
- Sullivan, K., & Lindgren, E. (2006). *Computer key-stroke logging and writing: Methods and applications*. Elsevier.
- Tate, T. P., & Warschauer, M. (2019). Keypresses and mouse clicks: Analysis of the first national computer-based writing assessment. *Technology, Knowledge and Learning, 24*, 523–543.
- Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing* (pp. 107–130). Oxford, UK: Elsevier.
- Yan, C. M. W., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A. M., & Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing, 25*, 1499–1521.
- Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice, 38*(2), 14–26.
- Zhang, M., Deane, P., Feng, G., & Guo, H. (2019). *Investigating an approach to evaluating keyboarding fluency*. Paper presented at the 2019 Society for Text and Discourse (ST&D) Conference, New York, NY.
- Zhang, M., Hao, J., Li, C., & Deane, P. (2018). Defining personalized writing burst measures of translation using keystroke logs. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th international conference on educational data mining* (pp. 549–552).
- Zhang, M., Zhu, M., Deane, P., & Guo, H. (2019). Analyzing editing behaviors in writing using keystroke logs. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzalez, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the Psychometric Society*. New York: Springer.

## 附录：写作过程指标、成分特征和效价

写作过程指标	成分特征效价	成分特征描述
单词间速度	+	单词间敲击空格键的速率中位数 (字符/s)
	-	单词间敲击空格键的平均持续时间 (单位ms, log转换值)
	-	单词间敲击空格键间隔的持续时间中位数 (单位ms, log转换值)
	-	单词间敲击空格键间隔的最小持续时间 (单位ms, log转换值)
	+	单词间附加击键 (包括空格和标点符号) 的速率平均数 (字符/s)
	+	单词间附加击键 (包括空格和标点符号) 的速率中位数 (字符/s)
	-	单词间附加击键间隔的持续时间平均数 (单位ms, log转换值)
	-	单词间附加击键间隔的持续时间中位数 (单位ms, log转换值)
	+	纯附加录入的连击长度中位数, 用600ms作为临界值来界定连击的断开 (字符数)
大连击流畅性	-	文本产出过程中错误 (后来纠正的打字错误或最终的拼写错误) 发生的相对概率, logit量尺
	-	最终文本中某一单词曾被错误拼写的相对概率, logit量尺
	+	纯附加录入的连击的平均数, 用8000ms作为临界值来界定连击的断开 (字符数)
	+	纯附加录入的连击的中位数, 用8000ms作为临界值来界定连击的断开 (字符数)
	+	纯附加录入的连击的最大值, 用8000ms作为临界值来界定连击的断开 (字符数)
	+	纯附加录入的连击的最大值, 用30000ms作为临界值来界定连击的断开 (字符数)
	-	所有连击长度的平均数, 连击边界界定为高于键间停顿时间中位数的8个标准差之上 (字符数)
	+	纯附加录入的连击的最大值, 连击阈界限定为包含停顿、持续时长8000ms以上 (字符数)
	+	纯附加录入的连击的平均数, 用30000ms作为临界值来界定连击的断开 (字符数)
	+	纯附加录入的连击的中位数, 用30000ms作为临界值来界定连击的断开 (字符数)
+	纯附加录入的连击的方差, 用30000ms作为临界值来界定连击的断开 (字符数)	
产出性	-	粘贴事件用时与单词内事件用时比值的对数
	+	击键总数
	+	句间标点符号击键总数
	+	单词内事件总数
	+	附加性击键总数
	+	全部大连击长度的标准差, 连击边界界定为高于单词内停顿时间中位数8个标准差之上 (字符数)
	+	全部大连击长度的最大值, 连击边界界定为高于单词内停顿时间中位数8个标准差之上 (字符数)
	+	单词起首事件总数 (即空格之后的字母或数字类字符)
删除性编辑	+	远距离跳跃性编辑行为速率的标准差 (字符/s)
	+	远距离跳跃性编辑行为速率的最大值 (字符/s)
	+	剪切事件总数
	+	剪切事件用时总长
	+	剪切事件相对总长, logit量尺
	+	剪切用时与单词内用时比值的对数

写作过程 指标	成分特征 效价	成分特征描述
删除性编 辑	+	剪切击键的标准差 (单位ms, log转换值)
	+	剪切击键延迟最大值 (单位ms, log转换值)
	+	剪切击键效率最大值-字母数或击键数/单词
跳跃性编 辑	+	跳跃性编辑事件用时与单词内非跳跃性事件用时的比值的对数
	+	跳跃性编辑用时总长
	+	同一单词内跳跃性编辑事件用时总长
	+	远距离跳跃性编辑用时总长
	+	跳跃性编辑击键持续时间最大值 (单位ms, log转换值)
	+	同一单词内跳跃性编辑击键持续时间最大值 (单位ms, log转换值)
	+	同一单词内的跳跃性编辑的跳跃距离的平均数 (字符数)
	+	同一句子内跳跃距离的标准差 (字符数)
	+	同一句子内跳跃距离的最大值 (字符数)
	+	远距离跳跃性编辑击键持续时间的最大值 (单位ms, log转换值)
	+	远距离跳跃性编辑的平均距离 (字符数)
+	句子内标点击键持续时间最大值 (单位ms, log转换值)	
句间过渡 和退格	+	句间空格击键速率的标准差 (字符/s)
	+	句间空格击键速率的最大值 (字符/s)
	+	句间间歇击键速率的最大值 (字符/s)
	+	句间间歇删除击键的速率的最大值 (字符/s)
	-	连续退格事件持续时间的最小值 (单位ms, log转换值)
段落特征	+	换行击键总数
	+	换行击键相对用时, logit量尺
	+	换行击键持续时间标准差 (单位ms, log转换值)
	+	换行击键持续时间最大值 (单位ms, log转换值)