心理学探新 2019,Vol.39,No.1,64-69
PSYCHOLOGICAL EXPLORATION

全国高考英语试题的城乡项目功能差异分析*

关丹丹 乔 辉 陈 康 韩奕帆
(教育部考试中心,北京 100084)

摘 要:本研究主要目的是针对高考英语成绩存在的城乡差异,检验这种差异是否来源于试题在城乡上的项目功能差异。如果两个能力本来相同的考生群体在某一试题得分上表现出不同程度的差异,该试题就存在项目功能差异。研究采用试题标准化分数差法,利用 STDIF 软件逐一分析了 2016 年三套全国高考英语卷的客观题是否存在城乡上的项目功能差异,在确定客观题没有项目功能差异后,以客观题成绩为匹配变量,采用条件得分图法对书面表达题是否存在城乡上的项目功能差异进行了分析。研究结果显示,高考英语全国 I、II、III 卷均未发现城乡上的项目功能差异试题,即可以认为高考英语全国卷对城乡不同户籍考生都非常公平、公正,城乡考生在英语成绩上的差异并非题目的公平性所致。

关键词:高考英语;项目功能差异;城乡
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2019)01-0064-06

1 引言

"促进公平、科学选才"是《国务院关于深化考试制度招生改革的实施意见》中提出的深化考试招生制度改革的主要指导思想(国务院,2014)。试题的公平性是大规模教育考试中广泛关注的重要问题之一,直接关乎考试的公平与科学。如果一项测试存在公平性问题,那么分数的解释、做出的决定及其后果都将是无效的、不合理的,甚至是有害的(Allalouf & Abramzon,2008)。这里所谓的测量公平性检验,心理测量学上称之为试题的项目功能差异检验。考试的命题人期望考试题目对不同的考生子群体是公平的,然而,在试题命制和考试实施过程中,不可避免地会受到一些无关因素的影响,有些因素可能会对不同考生群体产生不同的影响,使得能力本来相同的考生在试题得分上表现出不同程度的差异,这种差异被解释为项目功能差异(Differential Item Function,简称 DIF),其定义是指具有相同能力水平的考生,若在某试题上的得分因为考生组别的不同而显著不同时,则该试题可能存在 DIF。具有 DIF 的试题可能对某一组别的考生不利而对其他组别的考生有利,从而有违考试的公平性原则(Camilli,2006)。1986 年,美国教育测验服务公司(ETS)在编制测验的过程中规定:必须分析试题的项目功能差异,并且在分析试题的常规过程中加入一个 DIF 指数(曾秀芹,孟庆茂,1999)。我国学者也指出,测试开发机构应该本身或委托其它独立研究机构或学者进行测量公平性检验,并把检验报告公之于众(李清华,2016)。

高考英语考试是我国最具权威的语言测试,每年考生近千万,规模远超很多世界著名的语言测试,应该也必须重视项目功能差异的研究与分析。我国幅员辽阔,城乡之间经济文化发展很不平衡,农村考生的英语成绩总体上是略低于城市考生的,除了教育水平等差距之外,是否有的试题利于城市考生而不利于农村考生呢?即统计学意义上高考英语试题是否存在城乡 DIF 呢?

检测项目是否存在项目功能差异,关键在于如何判定两组被试是否具有相同的能力,即确定匹配变量。匹配变量既可以是考试自身(内部)的观察分数或者潜在能力值,也可以是附加测试(外部)的观察分数或者潜在能力值。在以考试自身分数作为匹配变量时,由于可能混有 DIF 试题的得分信息,检验过程中需要将有 DIF 的试题逐一剔除,以净化量表分数(Ferne & Rupp,2007)。检测 DIF 的方法有很多,总体上,基于 IRT 和结构方程的检验方法比较繁琐,基于 CTT 的方法(如采用观察分数法)对试题 DIF 进行检测比较简单易行。观察分数法用考试总分表征学生的能力,假设实验组和参照组的考试总分,即能力相同的考生,在同一试题上的得分率没有统计上的差别。观察分数法按照技术方法不同可分为不同类型,如 MH 法、STND 法等(曾秀芹,孟庆茂,1999;余跃等,2016)。其中,基于观察分数法的

* **基金项目**:全国教育科学规划单位资助教育部规划课题《新高考改革背景下的高考质量评价研究》(FBB160604)阶段性研究成果之一。

两阶段 DIF 检验法，在检验过程中将有 DIF 的试题会逐一剔除，在大样本下能够减少一类错误，提高统计检验效力（Zenisky et al.，2003）。

在语言测试 DIF 研究的文献中，就国家而言，美国是开展 DIF 研究最多的国家；就语言技能而言，阅读理解的 DIF 研究最多；就题型而言，0/1 二级计分的题型 DIF 研究最多（Ferne & Rupp，2007）。其中口语、写作这种评价产出性技能的试题由于是多级计分，进行项目功能差异检验的研究较少。多级计分题目往往主观性强，更易受到文化和环境因素的影响从而导致 DIF（张龙，涂冬波，2015）。但目前国内外学界焦点主要集中在二级计分题 DIF 检测方法的理论研究和应用上，对多级记分题 DIF 研究涉及较少。以写作为例，其中一个重要的问题就是检验写作 DIF 时两个群体的能力水平匹配变量怎么选，另外同时还有双评或多评带来的评分员功能差异分析（DFF）问题（Elder et al.，2003）。已有研究认为，对于写作的 DIF 检验，可以用语言测试其他部分的得分作为匹配考生水平的变量，当然其前提假设就是考生在其他部分的成绩与写作成绩相关很高（Lee et al.，2005；Ferne & Rupp，2007）。

目前，国内文献几乎没有针对高考英语考试是否存在项目功能差异试题的相关研究。本研究主要目的是检验全国高考英语试题在城乡上的项目功能差异情况。考虑到高考试卷的结构，既包括大量选择题，也包括书面表达题，本文将对两类试题分别进行项目功能差异检验，以期搜集高考英语考试是否公平公正的证据，从而回答城乡考生在英语成绩上的差异性是否与试题本身有关。

2　对象与方法

2.1　被试与数据

研究从 2016 年使用全国高考英语 I、II、III 卷的二十多个省中随机选择了 3 个省。使用 I 卷的 A 省英语考生共 333115 人，城市考生有 107490 人，占比 32.3%；农村考生有 225625 人，占比 67.7%。使用 II 卷的 B 省英语考生共 176722 人，城市考生有 105518 人，占比 59.7%；农村考生有 71204 人，占比 40.3%。使用 III 卷的 C 省英语考生共 275394 人，城市考生有 80611 人，占比 29.3%；农村考生有 194783 人，占比 70.7%。

2.2　研究材料

研究选取 2016 年全国高考英语 I、II、III 卷试题的考试数据进行 DIF 检验，整套试卷由 81 题构成：包括 20 个听力试题、20 个阅读理解试题、20 个完形填空试题、10 个语法填空试题、10 个短文改错试题，以及 1 个书面表达试题。其中，前 80 个试题总分为 125 分，可以统一按照二级计分的客观题来处理；书面表达题为多级计分，满分为 25 分。

2.3　数据分析过程及方法

一是对城市考生与农村考生的英语成绩进行描述统计分析，并对总成绩是否存在城乡差异进行 t 检验。二是对考试分数的内部一致性信度，以及是否符合单维性假设进行检验。三是对有所试题是否存在 DIF 进行分析。

对于客观题的 DIF 分析采用试题标准化两阶段分数差法（Dorans & Holland，1993；Zenisky et al.，2003）。对于书面表达题的 DIF 分析选用条件 P 值法，考生能力的匹配以客观题总分为依据，将参照组和实验组的考生采用每 5 分一个点进行粗分层（thick matching slicing），划分成不同的能力水平单元，计算每个能力水平单元在书面表达试题上的平均得分或得分率，绘制成图，以直观地比较相同能力水平的城市考生和农村考生在书面表达题平均得分上的差异（Zenisky et al.，2004；杨志明，2017）。

数据分析采用 SPSS 20.0 和 STDIF（April et al.，2016）软件。STDIF 软件在检测试题 DIF 时可以提供两个指标，一个是有符号的 DIF 指数（signed DIF），简称 SDIF，SDIF 适用于检测具有一致性 DIF 的试题。其计算公式为：

$$SDIF = \sum_{s=0}^{K} w_s (p_s^R - p_s^F)$$

其中，K 表示考试的满分值，p_s^R 为考试总分为 s 的参照组所有考生的条件难度系数；p_s^F 为考试总分为 s 的实验组所有考生的条件难度系数。w_s 是每个分数等级标准化的权重。

另一个是无符号的 DIF 指数（unsigned DIF），简称 UDIF，UDIF 适用于检测非一致性 DIF 试题。其计算公式是：

$$UDIF = \delta \sum_{s=0}^{K} w_s \left| p_s^R - p_s^F \right|$$

δ 是提供试题 DIF 方向的系数，δ 为“+1”，则试题有利于参照组，为“-1”，δ 则试题有利于实验组（April et al. 2016）。

3　结果

3.1　城乡考生考试成绩差异分析

城乡考生的英语考试成绩的平均分、标准差及两个群体的差异如下：

表 1 高考英语城乡考生考试成绩差异分析

	城市		农村		差异	效果量
	平均数	标准差	平均数	标准差		
I 卷						
总分	92.76	31.35	88.29	28.86	4.47	0.15
听力	22.70	5.35	21.28	5.03	1.42	0.28
阅读理解	26.28	9.94	25.20	9.44	1.08	0.11
完型填空	14.66	6.79	13.71	6.20	0.95	0.15
语法填空	8.72	4.06	8.39	3.83	0.33	0.08
短文改错	4.66	3.24	4.55	3.05	0.11	0.04
书面表达	15.73	5.62	15.16	5.47	0.57	0.10
II 卷						
总分	78.28	28.34	69.90	24.64	8.38	0.32
听力	12.30	8.01	10.56	6.41	1.74	0.25
阅读理解	23.72	8.11	21.91	7.39	1.81	0.24
完型填空	16.25	6.59	14.74	5.99	1.51	0.24
语法填空	7.27	4.44	6.28	4.18	0.99	0.23
短文改错	4.46	2.97	3.85	2.78	0.61	0.21
书面表达	14.26	5.89	12.54	5.93	1.72	0.29
III 卷						
总分	83.95	35.55	74.04	33.39	9.91	0.29
听力	19.95	6.74	17.59	6.43	2.36	0.36
阅读理解	24.15	9.36	21.81	8.87	2.34	0.26
完型填空	16.36	7.19	14.84	6.86	1.52	0.22
语法填空	6.36	4.35	5.51	4.05	0.85	0.20
短文改错	3.79	3.16	3.16	2.90	0.63	0.21
书面表达	13.34	7.81	11.13	7.86	2.21	0.28

总的来看，城市考生的英语成绩无论是总成绩还是各分项成绩，均高于农村考生。经检验 I、II、III 卷总分的城乡差异均显著（$p < 0.001$），城市考生的英语成绩平均比农村考生高出 4 ~ 10 分不等，因省份和所使用卷种而有所不同，特别是使用 II 卷的 B 省和 III 卷的 C 省，各项差异的效果量均大于 0.20，显示城乡考生差异较大。据此，是否可以说城市考生的英语能力水平明显高于农村考生呢？还不能，因为分数的差异可能是能力水平差异的真实反映，也有可能是题目本身对某一类群体更为有利造成的。因此，必须对试题是否存在城乡 DIF 进行分析。

3.2 高考英语考试的内部一致性与单维性分析等

表 2 各英语卷种的 a 信度与单维性检验情况

卷种	内部一致性 a 系数		单维性检验		书面表达与其他部分的相关系数
	含书面表达	不含书面表达	第一特征值	第二特征值	
I 卷	0.93	0.95	18.53	2.43	0.77**
II 卷	0.90	0.93	14.62	3.02	0.71**
III 卷	0.92	0.95	20.65	2.75	0.86**

可见，各英语卷的克伦巴赫 a 系数均大于 0.90，内部一致性非常高。因素分析显示，各卷种的第一特征值均远远大于第二特征值，除 II 卷外，都在 5 倍以上，基本可以认为英语考试是单维的。由于试卷中除去书面表达后的内部一致性更高，对于除书面表达外的客观题进行 DIF 分析时宜使用客观题总分作为考生能力水平匹配变量。另外，书面表达与其他部分成绩的相关系数均在 0.70 以上，且在 0.001 水平显著，属于高相关。这说明，在检验书面表达题目是否存在 DIF 时，可以用其他部分的成绩作为考生能力水平的匹配变量（Ferne & Rupp, 2007）。

3.3 客观题的城乡 DIF 分析

以客观题总分为考生能力水平匹配依据，计算得出各个题目的 SDIF 和 UDIF 值，将其绘制成图，全国 I、II、III 卷客观题的城乡项目功能差异分析结

果见图 1 ~3。

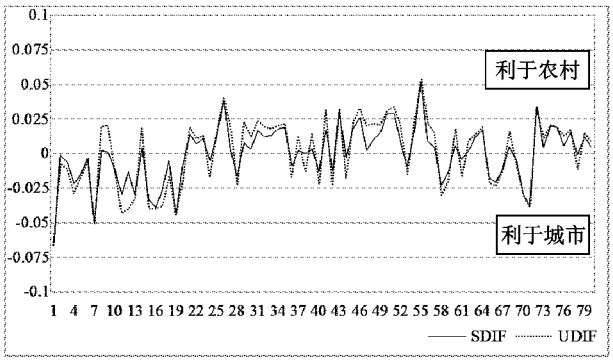

图 1　全国 I 卷英语试题城乡 DIF 指标

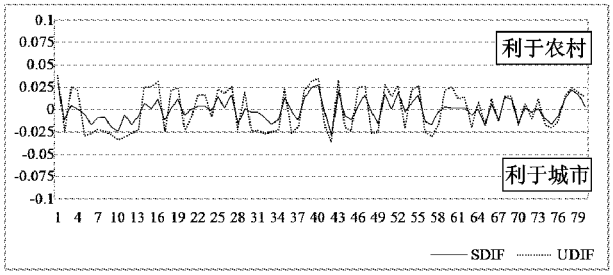

图 2　全国 II 卷英语试题城乡 DIF 指标

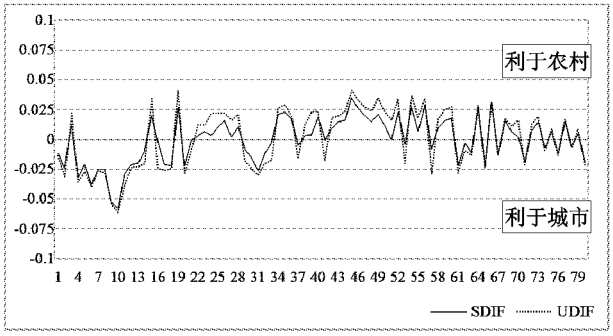

图 3　全国 III 卷英语试题城乡 DIF 指标

根据 STDIF 手册，如果 SDIF 或 UDIF 的绝对值小于 0.075，则表示该试题没有明显的功能差异；如果 SDIF 或 UDIF 的绝对值在 0.075 和 0.10 之间，则表示该试题有潜在的功能差异，但无需进行功能差异研究；如果 SDIF 或 UDIF 的绝对值大于 0.10，则表示试题具有明显的功能差异，需要进一步研究功能差异的来源（Zenisky et al.，2016）。该研究中 SDIF 和 UDIF 的值为正值，则有利于农村考生；若为负值，则有利于城市考生。上述检验结果显示，全国 3 套英语试卷中所有客观试题的 SDIF 和 UDIF 的绝对值都小于 0.075，试题不存在城乡的项目功能差异。

3.4　书面表达题的城乡 DIF 分析

由于三套试卷除书面表达试题外的所有题目都不存在城乡 DIF，可用除书面表达题外的成绩作为能力水平匹配依据，将考生从零分到最高分划分成不同的能力水平组，依据条件 P 值法绘图，比较相同能力水平的城市考生和农村考生在书面表达题上平均得分的差异。

全国 I、II、III 卷书面表达题的城乡项目功能差异分析结果见图 4 -6。

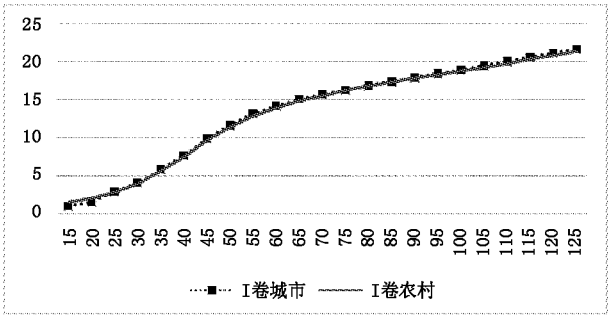

图 4　全国 I 卷英语书面表达试题得分城乡差异

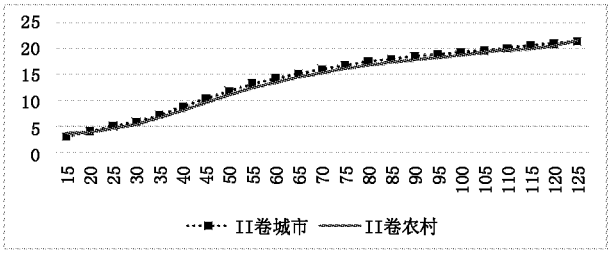

图 5　全国 II 卷英语书面表达试题得分城乡差异

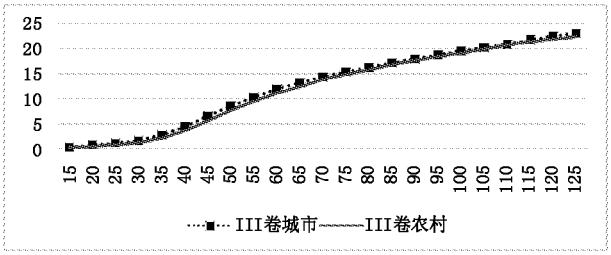

图 6　全国 III 卷英语书面表达试题得分城乡差异

从图形即可直观判断，相同能力水平的城市考生和农村考生在英语书面表达试题上的得分几乎是一致的，可以认为全国 I、II、III 卷的书面表达题都不存在城乡 DIF。

4　讨论与结论

从研究结果来看，2016 年高考英语全国 I、II、III 卷，无论是客观题，还是书面表达题，均不存在城乡 DIF 试题，即高考英语全国卷对城乡不同户籍考生都非常公平、公正。尽管城市考生与农村考生在英语成绩上是有明显差异的，但这种差异表现出的是两类考生英语水平的真实差异，并非题目的公平性所致。至于导致城乡两类考生英语水平差异的原因，可能与城乡经济条件、教学环境等有关，值得进

一步研究,但不在此文讨论范围。

在做 DIF 分析前,检验测验分数的内部一致性信度以及单维性等还是很有必要的,但却容易被研究者所忽视(Ferne & Rupp,2007)。该研究严格的进行了检验,确保了项目功能差异分析的前提条件是成立的。由于题目内容、题型都有可能会对某组考生有利,因此除了对客观题进行 DIF 分析外,研究者还尝试对书面表达题进行 DIF 分析。对于客观题的 DIF 分析在方法选取上主要考虑了高考英语试卷中客观题虽然都是 0/1 二级计分,但不同题型的最大分值不同,有每题 1 分,也有每题 2 分等多种情况。采用试题标准化两阶段分数差法,既可以通过两阶段的 DIF 分析将有 DIF 的试题逐一剔除,又可以很好的解决同是二级计分但最大分值不同的情况(Dorans & Holland,1993;Zenisky et al.,2003)。对于书面表达题的 DIF 分析,常用的多级计分题目的 DIF 分析有非参数检验法和参数检验法,以基于 IRT 为代表的参数检验法操作上比较复杂,结果不易于理解(张龙,涂冬波,2015)。而且与国际上许多知名的英语考试写作评分等级设定为 5 ~ 9 个相比,高考英语评分等级多达 26 个,数据结构与国外标准化考试存在明显差异。因此,已有的统计方法不一定适用,研究者依据 DIF 的定义,在通过内部一致性分析和因素分析确保了用客观题成绩作为能力匹配指标是可信和有效的情况下(Buzick & Stone,2017),以条件得分图形的形式直观展示相同能力水平的城市考生和农村考生在书面表达题平均得分上的差异,操作简单、结果容易理解,又非常有利于发现非一致性 DIF(杨志明,2017)。关于书面表达的 DIF 分析在国内外都是比较少见的,一方面是因为能力参照的匹配变量不好找,另一方面就是评分过程还会带来评分误差甚至是评分员的功能差异。该研究中三套试卷的内部一致性信度分析就显示,将书面表达题包含在内的 a 系数均是比不包含略有下降,可能就是受书面表达题有一定的评分误差的影响。这也从另一个角度说明,书面表达题的项目功能差异是值得从多个角度深入研究的。另外,研究采用户籍类型作为区分城乡的分类变量,可能混有借读考生,在有条件的情况下使用学校所在地作为城乡分类变量或许更为合适。

随着招生考试制度改革的进一步深化,不同考试试卷的公平和公正性越来越得到考试利益相关体的关注,对于 2016 年高考英语三套试卷在城乡 DIF 方面的检测体现了考试研究者对 DIF 研究的重视。总体上,目前国内关于试题 DIF 方面的研究还很少,除城乡以外的 DIF 研究也应引起关注;另外,我国高考和西方以选择题等客观性题目为主的考试形式不同,大量大分值的主观性试题如何进行 DIF 检验,也值得进一步探讨和研究。

参考文献

国务院.(2014).*国务院关于深化招生考试制度改革的实施意见*. http://www.gov.cn/zhengce/content/2014-09/04/content_9065.htm

李清华.(2016).语言测试的公平性检验框架.*现代外语*,*4*,549-561.

杨志明.(2017).考试公平性之题目及试卷功能差异探析.*教育测量与评价*,*9*,5-12.

余跃,杜文久,周娟,秦菊香.(2016).LP 方法及其与三种常用 DIF 检测方法的比较.*心理科学*,*39*(3),720-726.

张龙,涂冬波.(2015).多级计分题项目功能差异常用检测方法及比较.*江西师范大学学报(自然科学版)*,*39*(5),441-448.

曾秀芹,孟庆茂.(1999).项目功能差异及其检测方法.*心理学动态*,*7*(2),41-47.

Allalouf,A.,& Abramzon,A.(2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly*,*5*(2),120-141.

Buzick,H.,& Stone,E.(2017). *Recommendations for Conducting Differential Item Functioning(DIF) Analyses for Students with Disabilities Based on Previous DIF Studies*[R/OL]. [2017-08-11]. http://www.ets.org/Media/Research/pdf/RR-11-34.pdf.

Camilli,G.(2006). Test fairness. In R. L. Linn(Ed.),*Educational measurement*(4th ed.,pp.220-256). Westport,CT:American Council on Education.

Dorans,N. J.,& Holland,P. W.(1993). DIF detection and description:Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer(Eds.),*Differential item functioning*(pp.35-66). Hillsdale,NJ:Lawrence Erlbaum.

Ferne,T.,& Rupp,A.(2007). A Synthesis of 15 Years of Research on DIF in Language Testing:Methodological Advances,Challenges,and Recommendations. *Language Assessment Quarterly*,*4*(2),113-148.

Lee,Y-W.,Breland,H.,& Muraki,E.(2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing*,*5*,131-158.

Zenisky,A. L.,Hambleton,R. K.,& Robin,F.(2003). Detection of differential item functioning in largescale state tests:A

study evaluating a two – stage approach. *Educational and Psychological Measurement*,*63*,51 – 64.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF Detection and Interpretation in Large – Scale Science Assessments: Informing Item Writing Practices. *Educational Assessment*,*9*(1 – 2),61 – 78.

Zenisky, A. L., Robin, F., & Hambleton, R. K. (2016). *Differential Item Functioning Analyses with STDIF: User's Guide* [Version 6/15/2009]. Available from: Ronald K Hambleton.

A Study on the Urban/Rural DIF Evaluation of the NMET

Guan Dandan　Qiao Hui　Chen Kang　Han Yifan

(National Educational Examinations Authority, Beijing 100084)

Abstract: The purpose of this study is to analyze test fairness of the 2016 National Matriculation English Test (NMET) through conducting differential item function (DIF) analyses. If the responses of two groups of students with the same level of language ability differ on a common item, then the item owns DIF values, which means the item poses different level of difficulty for the two groups. The descriptive statistics indicated that there was significant difference in NMET scores between urban test – takers and rural test – takers. The standardization approach was applied to assess the three NMET papers focusing on urban/rural test – takers by using STDIF software which can detect both uniform DIF and non – uniform DIF. DIF is also investigated for a single writing item using the conditional P – value method. The result shows that no DIF values were found in the three NMET papers between urban and rural test – takers, suggesting that the score difference between the groups could not be attributed to DIF.

Key words: National Matriculation English Test; differential item function; urban/rural