基于Office Open XML技术的机考作弊检测方法探究
——以全国计算机等级考试为例

杨　英　黄啸波
（教育部考试中心，北京 100084）

摘要： 基于真实环境的MS Office计算机化考试提升了考试的友好性，但也带来基于文件复制或内容复制的考试作弊安全隐患，需要研究新方法对此类作弊进行检测。通过分析Office Open XML文档部件中的相关参数，提出复制类作弊的检测方法，并以全国计算机等级考试的Word试题为例进行检验。研究结果表明，该方法能直接、有效地判断基于文件复制的作弊，对大部分基于内容复制的作弊也能进行有效判断，但也存在个别无法判断的情况，需要结合其他参数综合分析。

关键词： 基于计算机的考试；计算机等级考试；作弊检测；考试作弊；考试安全

【中图分类号】 G405　　**【文献标识码】** A　　**【文章编号】** 1005-8427(2020)11-0042-6

DOI: 10.19360/j.cnki.11-3303/g4.2020.11.008

国内外使用MS Office考试的项目有很多，如微软办公软件国际认证（Microsoft Office Specialist, MOS）、国际计算机使用执照认证（International Computer Driving Licence, ICDL）、全国计算机等级考试（National Computer Rank Examination, NCRE）等，这些项目都是计算机化考试，允许考生在计算机上直接启用本地安装的MS Office软件进行作答。基于真实环境的、开放式的考试形式增加了用户友好性，提升了考试体验，但也带来考试安全隐患，催生出新的作弊方式——基于文件复制或内容复制的作弊。这种作弊不同于传统纸笔考试的抄袭，其方式更隐蔽，很难在考试过程中被发现。

为找出考试过程中未被发现的作弊考生，一些考试项目通过考后数据分析进行检测，如司法考试、职称外语考试、公务员考试等大规模高利害考试均采取这一措施。这种措施主要是通过对考生成绩数据进行统计分析，进而找出作弊考生。由于大多数统计检测指标都是基于一定的概率标准，不是所有作弊考生都能通过统计指标检测出来，也并非所有被检测出的考生都是作弊者[1]；因此在实际应用中，应谨慎对待和解释统计学指标。为此，本研究尝试基于Office Open XML的格式文档，通过挖掘文档部件内部相关参数检测复制类作弊，并以NCRE的Word操作题为例检验该方法的有效性。

1　理论基础

1.1　Office Open XML简述

Office Open XML（以下简称“OOXML”）是微软公司制定的一种基于XML的文档存储格式，于2008年通过国际标准组织（ISO）认定，成为文档存储的国际标准。OOXML的核心是使用XML参考框架和标准ZIP格式，将一系列相关的文档部件压缩

收稿日期： 2020-03-27　　**修回日期：** 2020-07-13

基金项目： 国家教育考试科研规划2019年度课题“无纸化考试的公平性研究——以全国计算机等级考试为例”（GJK2019025）

作者简介： 杨　英（1981— ），男，教育部考试中心，助理研究员；
黄啸波（1975— ），男，教育部考试中心，助理研究员。

存储到一个被称为“文档包”的容器中[2]。文档包中有文档部件、部件关系和部件类型,其中:文档部件记录文档的数据和属性等信息,包括文档组成部件和文档属性部件;部件关系记录文档部件之间及文档部件与外部资源之间的关系信息;部件类型记录文档包内所有部件的类型信息。上述文件都通过标准ZIP方式保存在文档包中[3]。

自2007版MS Office被广泛使用以来,Word、Excel、Powerpoint都使用OOXML文档格式代替原来二进制格式进行文档存储。为方便理解,可将MS Office文档想象为一个压缩包,该压缩包的后缀名不是“.zip”,而是“.docx”“.xlsx”和“.pptx”等。将Word、Excel、Powerpoint文档的后缀名修改为“.zip”,然后用解压工具打开,其内部结构见表1。

表1　MS Office 2013版文档内部结构

文档类型	文档组成部件	文档属性部件	部件关系	部件类型
Word	word	docProps	_rels	[Content_Types].xml
Excel	xl	docProps	_rels	[Content_Types].xml
Powerpoint	ppt	docProps	_rels	[Content_Types].xml

以Word文档“素材.docx”为例,其内部结构见图1。其中:文档属性部件存储在“docProps”文件夹内,包含“core.xml”和“app.xml”2个文件;文档组成部件存储在“word”文件夹内,包含“document.xml”“settings.xml”等多个文件;部件关系和部件类型分别存储在“素材”根目录下的“_rels”文件夹和“[Content_Types].xml”文件中。

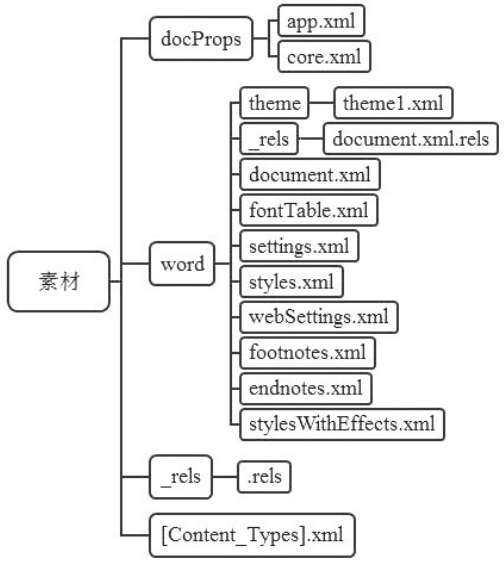

图1　Word文档内部结构图

1.2　关键参数

本研究关注2个重要参数:一是“settings.xml”文档部件中的“原始文档修订保存ID”(original document revision save ID,以下简称“risdRoot”),二是“document.xml”文档部件中的“段落字符格式修订ID”(revision identifier for paragraph glyph formatting,以下简称“rsidRPr”)。一份Word文档从创建到最终保存,其间的每次编辑保存都会产生一个唯一的“单次会话修订保存ID”(single session revision save ID,以下简称“rsid”),该值由8位十六进制数字组成(如00A46D61),与具体的编辑操作无关,只是用来表明本次编辑操作所在的编辑会话。一份Word文档产生的全部rsid值都保存在“settings.xml”文件的“修订保存ID列表”(listing of all revision save ID values,以下简称“rsids”)中,其中首次编辑保存ID即为原始文档修订保存ID,用rsidRoot表示。若2份Word文档的risdRoot值相同,表明这2份文档具有相同的首次编辑会话,即这2份文档的源头来自同一份文档[4]。表2中的2份Word文档,“文档_1”被编辑保存1次,“文档_2”被编辑保存2次。尽管它们是彼此独立的文档,但由于其具有相同的rsidRoot值“00256A13”,表明它们的源头来自同一份文档[5]。

表2　Word文档修订保存ID列表内容

文档_1	文档_2
<w:rsids> <w:rsidRoot w:val="00256A13"/> <w:rsid w:val="003B5A69" /> <w:rsid w:val="0049C65" /> <w:rsid w:val="00256A13" /> </w:rsids>	<w:rsids> <w:rsidRoot w:val="00256A13"/> <w:rsid w:val="003B5A69" /> <w:rsid w:val="0049C65" /> <w:rsid w:val="00256A13" /> <w:rsid w:val="00E98773" /> </w:rsids>

对Word文档中的文字进行编辑修改时，如果字符格式发生变化，将会生成一个唯一标识符记录本次操作，该值存储在所修改字符对应的运行模块中，用rsidRPr记录[4]。如将“素材.docx”中的“内容”2个字的颜色由黑色改为红色，则在“document.xml”部件中“内容”2个字所在的运行模块会增加新的rsidRPr记录。如图2中“内容”2个字所在运行模块增加了值为00B8392A的rsidRPr记录。

```
<w:p w:rsidRDefault="004C4DA6" w:rsidR="00EB659D">
  - <w:r>
      - <w:rPr>
            <w:rFonts w:hint="eastAsia"/>
        </w:rPr>
        <w:t>素材</w:t>
    </w:r>
    <w:bookmarkStart w:name="_GoBack" w:id="0"/>
  - <w:r w:rsidRPr="00B8392A">
      - <w:rPr>
            <w:rFonts w:hint="eastAsia"/>
            <w:color w:val="FF0000"/>
        </w:rPr>
        <w:t>内容</w:t>
    </w:r>
    <w:bookmarkEnd w:id="0"/>
</w:p>
```

图2 document.xml文档部件中rsidRPr值的截选

对Word文档进行“文件复制”“内容复制”“另存为”等操作时，rsidRPr值不会改变，除非再次对该部分文字的格式进行编辑修改[6]。根据上述特性，可使用rsidRPr值进行文档同源判断，如2个文档含有相同的rsidRPr值，说明这2个文档同源[7]。

1.3 相关研究

OOXML文档格式改变了人们处理数据的方式。当需要提取文档的文本内容时，可以只打开“document.xml”文档部件，而不必查看格式、属性等其他辅助部件[2]。根据该特点可进行Word文档的数据恢复，且文档内部一系列rsid值的生成原理和特性为Word文档溯源的司法鉴定和信息隐藏等提供了可行性。如林翔根据rsid生成的随机性、与文档内容无关的特性，通过修改“document.xml”中rsid值中的运行修订标识符（revision identifier for run，以下简称“rsidR”）实现对文字信息的隐藏，并能保证载体文档正常显示[8]。只要rsidR所在的文档段落未被全部删除，只对文档段内内容进行部分删除和修改并不影响隐藏的信息。付章杰等根据rsid值的唯一性判断可疑文档的来源，从而实现对Word复制文档的鉴别取证，并完成司法鉴定[9]。这种判断方式适用于所有采用OOXML格式存储的文档。

2 研究设计与实验

本研究使用rsid数据隐藏技术，在考试前对rsidRoot参数和rsidRPr参数进行修改，使其区别于历史值。通过rsidRoot参数检测考生作答文件与素材文件是否同源，进而判断考生是否存在文件复制的作弊行为。当检测到同源时，根据rsidRPr参数值不受内容复制等操作影响的特性，进一步检测其是否存在内容复制的作弊行为。具体实施分4步：试题素材预处理、考生模拟作答、参数提取及检测、检测结果及分析。

2.1 试题素材预处理

为模拟实际考试情况，以教育部考试中心编制的NCRE《计算机基础及MS Office应用》教程中一道Word练习题为例进行相关实验，该题要求对“文档2.docx”中的文字进行编辑、排版和保存[10]。

素材预处理分为5步：1）将“文档2.docx”转换为ZIP格式，打开“setting.xml”和“document.xml”部件，提取对应的rsidRoot值和rsidRPr值分别为“008E3E8D”和“00B605D0”；2）将上述rsid值分别添加到对应的rsidRoot和rsidRPr历史汇总库中；3）将“setting.xml”部件中的rsidRoot值和“document.xml”部件中所有的rsidRPr值都修改为“00EACE56”；4）将“setting.xml”和“document.xml”的修改时间恢复为“1980/1/1 0:00”，与其他部件一致；5）将“.zip”格式改为“.docx”格式，得到预处理后的素材文件。为避免考生发现素材文件的差异性，预处理后的素材文件仍命名为“文档2.docx”（为便于区分，以下统称“文档2-1.docx”）。

2.2 考生模拟作答

随机抽取NCRE考点10名一级MS Office科目考

生为研究对象。将前述试题题干、预处理后的素材文件“文档2-1.docx”和根据原始素材“文档2.docx”作答得到的标准答案提供给10名考生进行模拟作答,要求考生将作答结果统一命名为“文档.docx”后再提交。考生根据自己意愿,可以基于素材文件独立作答,标记为“正常作答考生”;也可以直接使用答案文件作答,标记为“文件复制作弊考生”;还可以复制答案文件的内容到素材文件内,标记为“内容复制作弊考生”。

2.3　参数提取及检测

文件作弊的考生通常是将提前准备好的答案文件复制到考生文件夹内,且该答案文件一般都是基于原始素材进行作答的结果,即答案文件与原始素材是同源的,具有相同的rsidRoot值。由于已经提前对素材文件进行预处理,修改了rsidRoot值,如果考生提交答案文件的rsidRoot值与原始素材的一致,可认定其为文件复制作弊。此外,除直接复制文件外,也有考生将答案文件打开后,将内容复制粘贴到素材文件内,即内容复制作弊,此时考生提交的作答文件与预处理后的素材文件具有相同的rsidRoot值。为找出内容复制作弊考生,可通过rsidRPr值进行判断,rsidRPr值不因内容复制操作而改变。素材文件往往包含多个文字段落,而题目本身一般不会要求对所有段落内的字符格式都进行修改,考生提交的答案文件中一般会留有素材文件中的某个rsidRPr值,如果考生提交的答案文件存在与原始素材文件中相同的rsidRPr值,就可以认定其为内容复制作弊。具体检测逻辑见图3。

根据考生提交的作答文件,提取其rsidRoot值和rsidRPr值集合,并分别与对应的预制值和历史库进行比较,结果见表3。其中rsidRPr提取的是“声音的强度与人体感受之间的关系”段落处对应的rsidRPr。根据题干的相关描述,此段文字不用进行字符格式的修改,考生正常作答后,该处的rsidRPr值应与所使用素材文件中的值一致,故可以用来进行与预设值和历史值的一致性判断。如该值发生变化,既不是预设值,也不是历史值时,就需要提取“document.xml”部件中所有rsidRPr值的集合进行判断。

2.4　检测结果及分析

根据作弊检测判断逻辑,10名模拟作答考生中有3人为正常作答,3人为文件复制作弊,2人为内

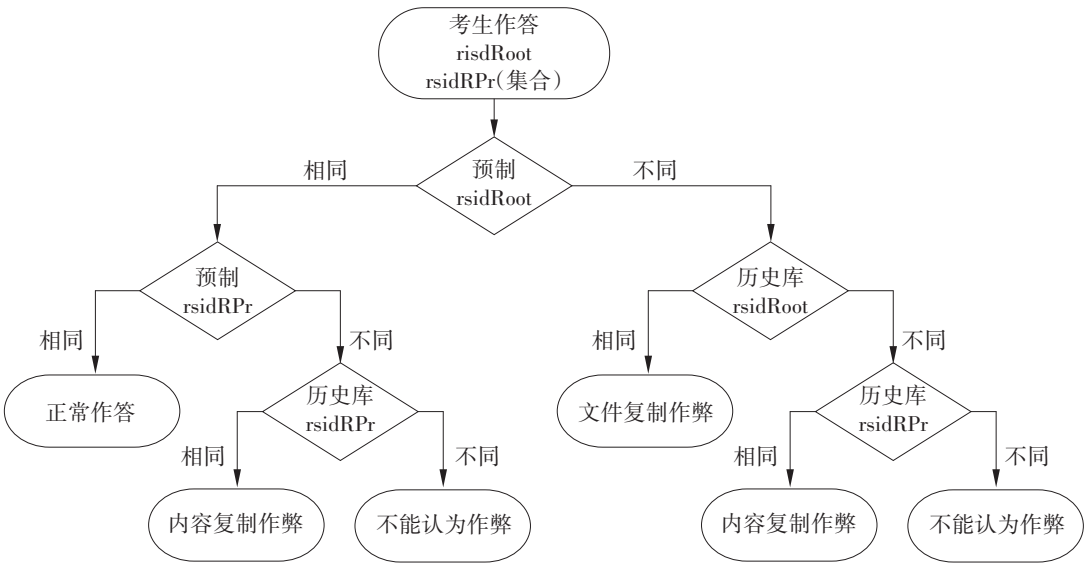

图3　作弊检测逻辑

表3 考生作答参数及检测结果

考生序号	rsidRoot	rsidRPr	预制rsidRoot	历史rsidRoot	预制rsidRPr	历史rsidRPr	检测结果
1	00EAEC56	00B605D0	相同	不同	不同	相同	内容复制
2	008E3E8D	00B605D0	不同	相同	不同	相同	文件复制
3	008E3E8D	00B605D0	不同	相同	不同	相同	文件复制
4	00EAEC56	00EAEC56	相同	不同	相同	不同	正常作答
5	00EAEC56	00EAEC56	相同	不同	相同	不同	正常作答
6	00864C5E	003B2F08	不同	不同	不同	不同	无法判断
7	008E3E8D	00B605D0	不同	相同	不同	相同	文件复制
8	00BB06E0	00B605D0	不同	不同	不同	相同	内容复制
9	00EAEC56	0067AC56	相同	不同	不同	不同	无法判断
10	00EAEC56	00EAEC56	相同	不同	相同	不同	正常作答

容复制作弊，其结果与考生实际操作一致。需要注意的是，有2类情况不能认定为作弊，但与正常作答也有区别。第一类情况是：rsidRoot值与预设值一致，但rsidRPr值与预设值和历史值都不一致，如考生9，其原因是考生未按照题干要求作答，对不用编辑的地方也进行了修改，导致rsidRPr值与预设值和历史值都不一致。第二类情况是：rsidRoot值和rsidRPr值与预设值和历史值都不一致，如考生6，这类情况也是因考生未按照题干要求作答引起的。考生未按题干要求基于原始素材作答，而是新建一个空白的Word文档，然后将答案文件或处理后素材文件的内容复制到该文档，并对所有字符格式进行编辑修改，进而导致rsidRoot值和rsidRPr值与预设值和历史值都不一致。

rsidRoot值检测和rsidRPr值检测各有优缺点。rsidRoot值不能检测内容复制作弊，但该值不受考生对文档内容编辑的影响；rsidRPr值可以检测内容复制作弊，但该值可能会受到考生对文档内容编辑的影响。实际操作中，可以将二者结合起来进行分析，具体流程见图4。当考生作答文件的rsidRoot值与素材中的预设值不一致时，表明考生不是根据素材文件进行的作答；当考生作答文件的rsidRoot值与历史值一致时，无须再检测rsidRPr值，可直接认定为文件复制作弊；当考生作答文件的rsidRoot值与历史值不一致时，需通过rsidRPr值检测是否存在内容复制作弊。

3 研究结论与讨论

计算机化考试越来越普遍，但其背后的问题往往被表象掩盖，被人们忽视[11]。与纸笔考试相比，基于计算机的考试作弊方式更加多样，手段更加隐蔽。传统的基于考生作答反应的统计学作弊检测方法，检测对象多为考生之间的雷同作答，即抄袭[12]，其检测题型只适用于选择题，对其他题型并不适用[1]；因此需要研究新方法进行无纸化考试操作题的作弊检测。

本研究依据OOXML格式文档中rsidRoot值和rsidRPr值的特性，提出针对MS Office操作题的作弊检测方法。该方法基于电子文档的证据进行判断，既可以作为附加证据，也可以直接用于作弊检测；

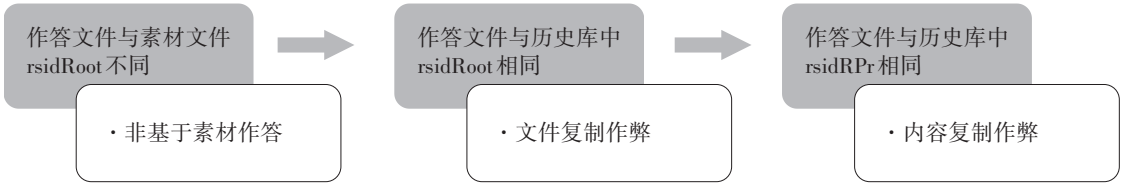

图4 作弊检测操作流程

与统计学作弊检测方法相比,能够更加直接、准确、有效地发现作弊行为。然而,该方法也有不足之处,如内容复制作弊检测会受考生操作影响,进而导致个别情况无法判断,需要结合其他参数综合分析,或通过增加系统录屏等功能采集更多的证据信息,从而进一步完善作弊检测,维护考试公平。

参考文献

[1] 黄宁. 答案抄袭指标和个人拟合指标对于不同作弊情境检测效力的比较研究[D]. 北京: 北京师范大学, 2009.

[2] RICE F. Introducing the Office (2007) Open XML file formats[EB/OL]. [2019-12-26]. https://docs.microsoft.com/en-us/previous-versions/office/developer/office-2007/aa338205(v=office.12).

[3] ECMA. Standard ECMA-376: Office Open XML File Formats: ECMA-376: 2016[S/OL]. [2019-12-18]. http://www.ecma-international.org/publications/standards/Ecma-376.htm.

[4] International Organization for Standardization. Information technology—Document description and processing languages—Office Open XML File Formats—Part 1: Fundamentals and Markup Language Reference: ISO/IEC 29500-1: 2016[S/OL]. [2019-12-19]. https://www.iso.org/standard/71691.html.

[5] Microsoft. DocumentFormat. OpenXML2.8.1[EB/OL]. [2019-11-05]. https://docs.microsoft.com/zh-cn/dotnet/api/documentformat.openxml.wordprocessing.rsid?view=openxml-2.8.1.

[6] 罗钊, 麦永浩. 基于RI码计算的Word复制文档鉴别[J]. 信息安全研究, 2016(4): 324-327.

[7] 罗文华, 孙道宁. Office Word文档溯源方法研究[J]. 警察技术, 2015(4): 45-47.

[8] 林翔. 反计算机取证之数据隐藏技术探究[J]. 电脑编程技巧与维护, 2016(13): 54-55.

[9] FU Z J, SUN X M, LIU Y L, et al. Forensic investigation of OOXML format documents[J]. Digital Investigation, 2011, 8(1): 48-55.

[10] 教育部考试中心. 全国计算机等级考试一级教程: 计算机基础及MS Office应用[M]. 北京: 高等教育出版社, 2019: 146.

[11] 赵宇鸣. 机考题库建设存在的问题和隐忧[J]. 现代企业教育, 2012(15): 84-85.

[12] 甘媛源. 考试抄袭识别的统计方法及应用研究[D]. 南京: 南京师范大学, 2012.

A Study on Computer-based Test Cheating Detection Based on Office Open XML File Formats

YANG Ying, HUANG Xiaobo

(National Education Examinations Authority, Beijing 100084, China)

Abstract: Based on real environment, the MS Office exam makes itself more user-friendly, but it also brings about the security risk of cheating on test in copying document or content. Therefore, new methods are needed to detect such kind of cheating. By analyzing the relevant parameters in the Office Open XML document components, this study proposed a detection method of copying cheating, and conducted a validation process by taking a Microsoft Word item from National Computer Rank Examination as an example. The result shows that this method can make direct and effective judgments on all document-copying cheating as well as the majority of content-copying cheating. For some individual cases that cannot be judged by this method, we need to make more comprehensive analysis together with other parameters.

Keywords: computer-based test; National Computer Rank Examination; cheating detection; cheating on test; test security

(责任编辑:张　丽)