

博士学位论文盲评规则的定量实证分析

李国辉¹，靳冬欢²，张宇航¹，武小悦¹，刘盼¹

(国防科学技术大学 1. 信息系统与管理学院; 2. 研究生院, 湖南 长沙 410073)

摘要：以博士学位论文盲评的评阅数据为分析对象，通过研究评阅专家数量、评阅规则与论文评阅风险之间的关系，对评阅过程中可能存在的论文评阅风险进行了研究，并重点分析了评阅专家数量、评阅规则的变化对风险值的定量影响。

关键词：博士学位论文评阅；盲评；风险分析

中图分类号：G643.8 **文献标志码：**A **文章编号：**1672-8874 (2015) 04-0078-05

A Quantitative Analysis of the Rules for Blind Review of Doctoral Dissertations

LI Guo-hui¹, JIN Dong-huan², ZHANG Yu-hang¹, WU Xiao-yue¹, LIU Pan¹

(1. College of Information System and Management, 2. Graduate School, National University of Defense Technology, Changsha 410073, China)

Abstract: Analyzing the data of blind review of doctoral dissertations, making a research of the relationship between the number of reviewing experts, reviewing rules and the risks in dissertation-reviewing, the thesis studied the risks that might exist in the process of reviewing and analyzed the quantitative impact of the number of reviewing experts and the change of reviewing rules on the risk value.

Key words: review of doctoral dissertations; blind review; analysis of risks

判断一个学生是否达到博士学位水平，对其博士论文质量的盲评是一种重要的手段。全国高校普遍采用了博士学位论文的双盲评阅^[1]。学位论文双盲评审是指将隐去导师姓名和研究生姓名的论文送到有相同和相似专业的高校，请国内同行专家进行评阅，对论文是不是达到答辩要求给出结论、提出修改意见和建议。在这种评审方式，研究生和评阅专家彼此都不了解对方的具体情况，故名为双盲审评。

但由于专家学术评价的主观性、专业水平局限、宽严要求以及责任心等的不同，必然存在一定的评审误判风险^[2]。按照现行的评阅制度对博士论文进行评阅，博士论文双盲评阅主要面临两类风险：一类是将合格的论文评审为不合格的论文，我们称之为 α 风险；另一类是将不合格的论

文评审为通过，我们称之为 β 风险。后者的风险效应更大。

如何制定科学合理的双盲评阅规则，需要从保证博士论文质量水平以及质量保证的经济和时间成本（包括评阅成本、博士生推迟答辩等）等多个方面进行综合权衡。为此，必须分析不同的评阅规则、论文质量水平与评阅风险之间的量化关系。

目前博士论文评阅中存在许多问题，王则温、施亚玲等对相关问题进行了思考和讨论^[3-4]。但是关于博士学位论文评审的相关研究尚处于定性或简单的描述性统计分析阶段^[5]，缺少对于评阅机制与不合格论文通过评审的风险之间的定量关系研究。评审专家人数以及评审通过规则的制定主要是根据各高校的情况和管理经验确定^[6-7]，缺少

充分的理论指导和定量决策依据。

本文基于统计检验的原理，基于一定的假设，建立了一种博士学位论文评阅风险的量化分析模型，表示双盲评阅风险与评阅专家数量、评阅通过规则之间的定量关系。然后通过运用该模型进行案例分析，得出不同假设情况下两类风险水平的变化规律。希望能够为改进和完善博士学位论文评审机制提供一些决策参考依据。

一、博士学位论文评阅风险建模

为了构建模型，本文提出如下假设：

- (1) 学位论文本身的真实质量水平只分为合格(1)与不合格(0)水平两个等级。
- (2) 各个专家的评审结果只分为合格与不合格两种情况。
- (3) 各个盲评专家对于学术论文的判断水平特征相同。每个专家在评分时独立、客观、公正，且对于给定质量等级的博士学位论文的成绩定量服从正态分布函数（此假设在描述性统计分析中得到验证）。专家评阅判断的水平用误判概率表示。

假设盲评规则为：若 N 个盲评专家中有大于或等于 d 个认为论文不合格，则论文不能通过本次评审。我们根据产品检验的风险分析模型，构建论文评阅的风险模型^[8]。

当论文真实质量水平为合格时，论文按盲评规则把合格论文误判为不合格的情况出现的概率，用 α 风险表示。

当论文真实质量水平为不合格时，论文按盲评规则把不合格论文被误判为合格的情况出现的 β 风险。

设 $I_p(N-d, d+1)$ 是参数为 $(N-d, d+1)$ 的 Beta 分布的分布函数在概率 p 的取值。 p_{10} 表示一位专家将合格论文(1)判为不合格(0)的概率， p_{01} 表示不合格论文(0)被判为合格(1)的概率。文献[2]最终给出了 α 和 β 风险的计算公式：

$$\alpha = I_{p_{10}}(d, N-d+1) \tag{1}$$

$$\beta = I_{p_{01}}(N-d+1, d) \tag{2}$$

公式(1)和(2)给出了两种类型的评阅风险、盲评专家数量 N 与论文质量评阅判决阈值 d 之间的关系。

图1示出了 N 分别为3, 5, 7, 9, d 取1~9时

两类风险水平变化的情况。 α 风险用实线表示， β 风险用虚线表示。

从图中看出以下规律：当盲评专家总数 N 一定时，随着评阅不合格数目阈值 d 的增加， α 风险变小，而 β 风险变大。因此，仅改变 d 不能同时降低两类风险水平。之间有一个权衡值。例如，只有1名专家判别不合格，那么就认为该论文不合格的风险相当大；多几位专家判别为不合格，判为不合格的风险降低。反之，只有1名专家判别为不合格，就认为该论文合格的风险较小；多几位专家判别不合格，认为该论文合格的风险就变大。因此，通过增加盲评专家数量和提高专家评阅判断水平，可以同时降低评阅机制的两类风险。特别是专家判断水平的提高将可使风险明显下降。

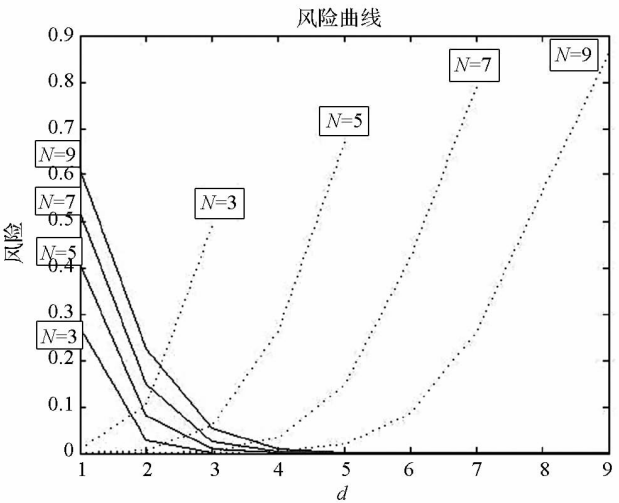


图1 两类风险水平的变化 ($p_{01} = 0.2, p_{10} = 0.1$)

二、描述性统计分析

我们抽取某大学2013年的博士学位论文盲评数据。把一位专家对一篇论文的评阅结果作为一条评阅记录。数据集中，共有351篇博士论文，2430条评阅记录，其中明评记录1570条，盲评记录860条。盲评记录中有77条复评记录，783条非复评记录。在评阅书中，针对论文水平的评价分为三个等级：达到博士论文水平、基本达到博士论文水平、尚未达到博士论文水平；针对是否答辩的评价也分为三个等级：准予答辩、修改后答辩、不予答辩。评价得分按照100分计算。

有2条记录，专家没有给论文打分，只给出了尚未达到博士论文要求的结论和不予答辩的建议。根据实际数据，在处理的过程中将这两条记录的

分值设为 50 分来进行处理。

一些专家给出的建议是准予答辩，结论却是基本达到博士学位论文要求；还有一些专家，其结论是达到博士学位论文要求，却建议修改后答辩。基本达到博士论文要求或修改后答辩的评价，意味着论文内容有一些缺陷。谨慎起见，我们把这些记录列入有异议的情况。为此，我们把给出的结论中包含基本达到或尚未达到博士学位论文要求，或答辩建议中是修改后答辩或不予答辩的

记录作为有异议记录。我们认为评判记录中只有同时满足“达到博士学位论文要求”和“准予答辩”的才算一个无异议的记录。

表 1 给出了评阅数据的描述性统计。分别统计评价数据的所有评价记录、明评记录、盲评记录、非复评的盲评记录、答辩建议为修改后答辩的盲评记录、有异议盲评记录、答辩建议为不予答辩的盲评记录。表中给出了这些记录的总数、均值、中值、众数、方差、偏度、峰度。

表 1 各类记录的分布特征

	所有	明评	盲评	盲评无异议	盲评有异议	盲评不予答辩
论文数量	2430	1570	860	684	176	7
均值	88.97	91.31	84.70	86.44	77.93	61.57
中值	90.20	91.20	85.95	87.00	79.00	61.00
众数	90.4	91	86	88	80	50
方差	27.361	7.880	34.702	17.178	45.460	100.952
偏度	-1.791	-0.868	-1.426	-0.682	-1.241	0.169
峰度	5.753	3.451	4.063	1.402	2.851	-1.374
极小值	50	70	50	66	50	50
极大值	99	99	99	99	93	76

由表 1 可以得到如下结果：

(1) 从各项分布来看，评阅分数的分布呈现正态分布。

(2) 明评记录的平均值(91.31)明显大于盲评记录的平均值(84.70)。去掉极值的影响，看中值(分别为 91.2 和 85.95)的取值，也反映了显著的差别。虽然达到明评标准的论文水平高一些，但是这个统计数据说明盲评专家给出分数比较客观。对于盲评，盲评无异议和有异议的均值和中值差别也比较明显。盲评有异议的均值和中值都低于 80 分，因此我们认为低于 80 分是一个警告阈值。

(3) 明评记录的方差(7.880)明显小于盲评记录的方差(34.702)。方差大表示评价的意见的分歧大，分数分布较分散，表明盲评专家敢于给出客观的评价。评价的显著差异主要体现在盲评有异议的论文上，方差达到 45.46。盲评无异议的评价差异并不太大，但是还是比明评差异大。

(4) 偏度是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征。偏度的值可以为正，可以为负或者甚至是无法定义。直观看来就是密度函数曲线尾部的相对长度。

偏度为负值，表明评价分数偏小的值较多。从表 1 统计数据看，盲评要比明评的偏度大，而盲评有异议又要比盲评无异议的偏度大，说明打低分的评价较多。从均值、中值和众数的取值看，明评的均值大于中值大于众数，说明专家评价偏重于打较高的分，这与实际评价吻合。而明评中，表现出众数大于中值大于平均值，说明评价打分偏高。

(5) 峰度表征概率密度分布曲线在平均值处峰值高低的特征数。直观看来，峰度反映了尾部的厚度，描述分布形态的陡缓程度。在相同的标准差下，峰度系数越大，分布就有更多的极端值，那么其余值必然要更加集中在众数周围，其分布必然就更加陡峭。从统计结果看，盲评比明评的峰度值高，盲评有异议比盲评无异议的峰度值高，表明后者在评价打分中集中在众数周围的值较多，在众数值周围的峰值较为集中。

(6) 盲评和明评中都有最高打分 99 分的，这是专家的主观打分，我们不予置评。但是我们注意到最低分的情况，明评为 70 分，盲评无异议的为 66 分，都比较低，但是都给予了达到博士论文

要求和准予答辩的评价。这种情况很少，但是要引起我们的注意。

三、评阅风险分析

我们抽取某大学 2013 年的 102 篇七位专家盲评（以下简称七盲评）的论文，共 $102 \times 7 = 714$ 个数据中，共有 150 条有异议记录。我们有如下的判断准则。

判断准则 1：考虑到专家将结论判为尚未达到博士学位要求或者将答辩建议判为不予答辩都是十分谨慎的，因此可以认为只要出现上述两种情况，无论评分是多少，就认为论文不合格。即，认为其他对论文无异议的专家判断错误，将不合格论文误判为合格。

判断准则 2：考虑到专家判有异议都是十分谨慎的，当七位专家有三位或者三位以上专家判断一篇论文有异议，这篇论文不合格的概率极大，因此可以认为是其他几位专家判断错误，将不合格论文误判为合格。

判断准则 3：对于有两条有异议记录的论文，若该条记录的专家对其打分都小于 80 分或有一个低于 70 分，则认为该论文不合格。则其他五位判断错误，将不合格论文误判为合格。

判断准则 4：对于有一条有异议记录的论文，若该条记录的专家对其打分小于 70 分，则认为该论文不合格，则其他六位专家判断错误，将不合格论文误判为合格。

根据以上四个判断准则，共有 27 篇为不合格论文被错误评定为合格论文，其中 1 篇有六条异议

记录，1 篇有五条异议记录，5 篇有四条异议记录，9 篇有三条异议记录，11 篇有两条异议记录。

这样，将不合格论文误判为合格的记录总数为： $1 \times 1 + 2 \times 1 + 3 \times 5 + 4 \times 9 + 5 \times 11 = 109$ 。

专家将不合格论文误判为合格的概率 $p_{01} = 109/714 = 0.1527$ 。

由于数据集中没有评价为合格又被抽查不合格的论文，因此无法估计专家将合格论文误判为不合格的概率 P_{10} 。正常情况下，专家将一篇论文判为不合格都是十分谨慎的，故而 p_{10} 都比较小。在博士论文评阅过程中，我们主要关注并控制 β 风险（将不合格论文评为合格论文的风险），而可以容忍一定的 α 风险。因而我们可以在达到我们对 β 风险的要求的前提下，尽量减少 α 风险，即在满足对 β 风险控制的前提下，将 d 值取尽量大。

图 2 是表示了 β 风险的计算结果，分别给出上限 0.001 和 0.0001 的风险曲线图。当把 β 风险上限设置不断降低，更多细节得到显示。在风险上限设置为 0.001（千分之一）的左图中，各点的风险都小于千分之一。在风险上限设置为 0.0001（万分之一）的右图中，各点的风险都小于万分之一。在七盲评过程中（即 $N = 7$ 时），根据左图中，当 $d = 3$ ，能够将 β 风险控制在千分之一点三左右，即盲评专家中有小于等于二的评语认为不合格，就让该论文通过的风险大约在千分之一左右，一千篇论文中有一篇不合格论文被放出去。在右图中表示当 $d = 2$ ，能够将 β 风险控制在万分之一以内，即盲评专家中有一个评语认为不合格，就让该论文通过的风险大约在万分之一左右，一万篇论文中有一篇不合格论文被放出去。

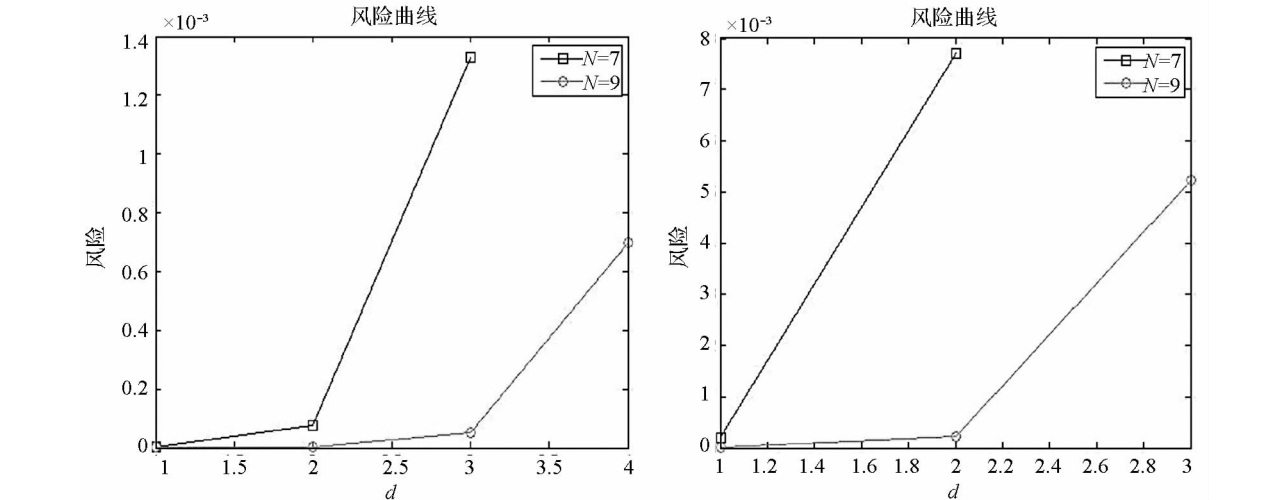


图 2 对 β 风险上限设置分别为 0.002 和 0.0001 的风险曲线图

四、总结与建议

本文对博士论文评论数据风险进行了分析研究。建立博士学位论文质量等级的概率分布模型,以及在不同等级下评阅成绩的概率分布模型,并根据实际数据进行了分析。通过计算分析,研究确定专家数量、专家水平、评审通过规则与论文评审风险水平之间的定量关系。各高校和研究机构,可以根据自身数据得到统计和经验性的专家犯错概率 p_{10} 和 p_{01} ,从而用本文的方法设置合理的评阅规则 N 和 d 。

参考文献:

- [1] 刘晓元,周珞晶,方毅.博士学位论文评阅制度改革设想[J].学位与研究生教育,2010(10):24-27.
- [2] 武小悦.博士学位论文双盲风险分析[J].高等教育

研究学报,2014,37(3):47-50

- [3] 王则温,赵张耀.关于博士学位论文评审有关问题的探讨[J].学位与研究生教育,2009(3):38-42.
- [4] 施亚玲,李若英,杜鹃,等.关于博士学位论文评价体系若干问题的思考[J].华南理工大学学报:社会科学版,2011,13(5):105-110.
- [5] 孙妍.谈博士学位论文隐名评审中非共识评阅意见的处理[J].中国高教研究,2004(10):40-42.
- [6] 周珞晶.关于我校博士学位论文评阅和答辩规定的思考[J].高等教育研究学报,2007(3):93-94.
- [7] 张宇航.博士学位论文评阅制度改革研究与实践——以国防科学技术大学为例[D].长沙:国防科学技术大学研究生院,2006.
- [8] 武小悦.博士学位论文双盲风险分析[J].高等教育研究学报,2014,37(3):47-50.

(责任编辑:胡志刚)

(上接第52页)

强调无线能量传输应用前景的案例:无线充电手机和药囊内窥镜。

由于该题目的研讨略微超出教材范围,因此特意在研讨前要求提前一天发给参与研讨交流的其他学员阅读。实践证明,对于“军事信息技术基础”这样综合性很强的课程来说,翻转课堂中教员适当的引导和放手可以充分发挥学员的主观能动性,同时以点带面拓宽学员的科学视野和格局,极大提高翻转课堂教学的效果。

四、结束语

MOOC 翻转教学是一种新型教学方式,它将学员推到教学活动主体地位,培养学员自主学习、积极思维的能力。从培养学员具有适应未来需要的潜能为目的出发,定位为军兵种课程的“军事信息技术与系统”MOOC 翻转课程教学对培养学员良好的思维习惯有着极其重要的作用,符合军队信息化建设的发展需求。本文以“军事信息技术与系统”MOOC 翻转课程为例,针对 MOOC 翻转教学实践中发现的一些问题,给出了相应的解决办法和对策,希望能对大学 MOOC 翻转课堂教学起到一定推进作用,从而整体提升教学效果和质量。毕竟 MOOC 翻转教学是一个新兴的教学方式,我们虽然有一定的教学实践体会,但是还有很多没有看清楚的地方和没有想明白的问题还需

要在今后的教学实践中积极探索,深入学习。

参考文献:

- [1] 吴剑平,赵可.大学的革命——MOOC 时代的高等教育[M].北京:清华大学出版社,2014:100.
- [2] 程吉宁.抓住机遇,迎接挑战,加速发展大规模在线教育[R].北京:清华大学“大规模在线教育论坛”主题报告,2013-06-03.
- [3] 孙茂松.MOOC 的基本技术特征及其发展趋势[R].广州:2013 年高等教育信息化创新论坛,2013-06-23.
- [4] 赵荣,马亮,张玉龙.MOOC 的理性思考:兴起、发展与未来[J].高等教育研究学报,2014(2):9-14.
- [5] 王广.合作教学与 MOOC 建设[J].高等教育研究学报,2015(2):46-51.
- [6] 王迎,刘惊铎,韩艳辉.MOOCs 在中国发展的理论思考与实践探索[J].中国电化教育,2014(1):52-60.
- [7] 徐明,龙军.基于 MOOC 理念的网络信息安全系列课程教学改革[J].高等教育研究学报,2013(3):16-19.
- [8] 李曼丽,张羽,叶赋桂.解码 MOOC——大规模在线开放教育的教育学考察[M].北京:清华大学出版社,2013:17.
- [9] 波利亚.怎样解题:数学思维的新方法[M].徐宏,冯承天,译.上海:上海科技教育出版社,2007:145.
- [10] 詹姆斯·格雷克.信息简史[M].高博,译.北京:人民邮电出版社,2013:11.
- [11] R·柯朗,H·罗宾.什么是数学:对思想和方法的基本研究[M].左平,张饴慈,译.上海:复旦大学出版社,2013:405.

(责任编辑:赵惠君)