

反思教育中的 考试、测试、分级和评分现象

约翰·怀特¹著,赵显通²译

(1. 伦敦大学学院 教育学院,英国 伦敦 WC1H0AL;2. 西南大学 教育学部,重庆 400715)

摘要:首先,探讨了考试、测试、分级和评分的概念以及它们之间的联系,阐明了两点内容:一是在教育领域中,所谓的考试并不是一项系统且彻底的调查(thorough investigation);二是评分的观念是如何成为考试的核心。而后,检视了自省(self-examination)的宗教渊源,以及考试上的成功与拥有美德之间的关联。最后,鉴于传统考试存在的诸多问题,提出了一种积极的替代方案,而此方案可以使学校的考试更加接近系统且彻底的调查。

关键词:考试;测试;分级和评分;自省;精英学校;形成性评价

中图分类号:G424.74 **文献标识码:**A **文章编号:**2095-8129(2017)05-0089-06

为何会存在学校考试制度?答案往往包含两个方面:一个是近期目标(immediate objective),另一个是长期目标(further objective)。很显然,考试是以一种公开记录(public record)的方式来说明一名学生在多大程度上理解和掌握了他所学科目的知识。在英国 A-level 考试中,取得优异成绩的学生比成绩差的学生无疑能更好地掌握这一领域的知识,无论是在深度还是广度上。

为什么这种公开的记录如此重要?首先,它能够帮助学生接受进一步的教育,尤其是进入高等教育领域继续学习,进而获取一份不错的工作。其次,考试成绩在其他方面也很重要,特别是在比较不同学校或是不同国家的教育质量方面能发挥作用。

上面所说的就是我们日常理解的学校考试的目的。初看之下,都是我们习以为常且司空见惯的看法。

一、何为考试

学校考试并非看上去那样简单。你越是仔细琢磨,就越能看出考试的“古怪”之处。前文提到过,考试可以提供一种公开的记录来说明一名学员在多大程度上理解了他所学的科目,我们大多数人都会同意这一观点。在此情况下,我们是在一个更加抽象的层面上表达对考试的理解。从广泛的意义上讲,考试是对某物的一种调查,并试图以一种系统(systematic)且彻底的(thorough)方式揭示某种现象的本质。正如某位哲学家对实用主义的研究,又如一名医生对一个疑似患有胆结石病人进行的医学检查。这些可检视的现象涵盖面非常广泛,从理论研究到人的身体疾患,再到法庭证据,可以说数不胜数。

显然,以上例子说明并不是所有的检查对象都是人,同样,并非一切针对人的调查都要考查人

收稿日期:2017-04-28

作者简介:约翰·怀特,英国教育哲学界领军人物,伦敦学派的代表人,伦敦大学学院教育学院教育哲学荣誉教授。

赵显通,哲学博士,西南大学教育学部讲师。

们了解多少学科知识。一位临床心理学家可以调查某位病人染上艾滋病之后的心理感受或是他对别人的态度等问题。若我们把考试限定在考查人们“知道什么”上面,那么它就不仅仅涉及调查学生掌握多少学科知识,也可以是别的方面。从广义的考试视角来看,学校考试存在一个问题。广义上的考试是对事物进行一种“系统且彻底”的调查,为了得到比较客观的解释,应采取多方位的视角。那么,学校考试能够达到此目标吗?据我所知,考官阅卷的时间通常很有限。比如:在 A-level 考试中,教师审阅一份关于 1815 年以来的欧洲历史的答卷,可能他所花费的时间不超过半小时,甚至更短,想要在如此短的时间内完成一项我们前文所述的“彻底的调查”是极为困难的,因为这种调查要看考生对所有主要欧洲国家 1815 年以来史实性知识的掌握程度,要看学生理解的深度和广度(包括经济、政治、军事、外交、社会以及其他因素),也要看他们是否考虑了 1815 年之前的重要历史事件。Andrew Davis 也阐述了为什么考生的这种联系性的理解(connected understanding)很难在一段简短的书面回答中体现出来^[1]。不同的考生思考的方向不同、关注的维度也不尽相同。传统的考官在审阅答卷时,无论如何也不可能对学生在这一领域的知识进行彻底的调查。如果他们确实想彻查,那么就要采用另外更加耗时的办法。例如:他们会看考生在某个话题上阅读了多少书籍,写过多少关于此话题的文章,并从教师或其他人那里获取证据,还会对考生进行长时间的访谈来考查他的理解有多广、多深。“彻底调查”意义上的考试并不是我们所熟知的考试。这也说明传统的学校考官也许并非总是(或曾经是)广义上的考官。但是,如果他们并未在考查,那在干些什么呢?

传统考试的工作就是分级(grading)和评分(marking)。分级并不一定包含着评分。比如:一位果农可以依据水果的大小进行分级,此处的分级意味着要对水果进行比较及排列,评分并不包括在内。但是,学校考试的分级是要打分,而后依据分数来评定是 C 级还是 B 级。严格地讲,分级和评分并不涉及考试。我们知道考官要对考卷打分评级,但这并不需要了解某人什么或者说进行彻底调查。也就是说,对某人的知识进行彻底的调查不一定要包括分级。正如我们并不给战俘打分评判他透露了军队的动向一样。

有没有一种既能彻底调查学生们了解什么,又能不牵扯到分级和打分的办法?我稍后再来讨论。虽说答卷获得的分数是一种客观的记录,但却受考官价值观的影响。在一种全国性的考试中有对学生理解力的全面调查,考官所给的分数会相差很大。即便所有的考官都有一套相同的价值观,也不一定能保证他们可以赋予每种价值观相同的权重。传统的解决方案是众所周知的,就是建立分级的体系,即用评级标准尽可能地消除这种不确定性。对于像现代语言学和数学这样的有着明确对错答案的科目来说,这个方案没什么问题,而其他科目则不然,考试内容很容易集中于知识点的考查而非对该科目更加深入的理解。

二、何为测试

前文讨论了考试,现在我们来看看何为测试(testing)。有些时候,我们会把考试和测试当作同义词。例如,我们可以把 GCSE^① 这样的考试当作“高风险测试制度”(high stakes testing regime)的一部分。但是,我们也会从一个更确切的意义上来探讨测试。

在对驾驶进行测试时,考官会检查学员是否知道如何驾驶汽车,但他不会对学员驾驶方面的理论知识进行一项“彻底的调查”,因而从严格的意义上讲,考官并不是在主持考试,他的判断在很大程度上依据的是感知性的证据。同样,如果想确认某人是否会写自己的名字或者是否会骑自行车,我们很大程度上依靠自己亲眼所见就可以得到答案。除了肢体技能之外,智力技能也类似。例如:我们可以毫不费力地看出一个孩子是否能拼写句子;我们可以通过儿童的计算看出他掌握加法和

① 英国普通中等教育证书——译者注。

减法的能力;借助其他的方式,也能知道他们是否可以进行一些初级的翻译工作。

测试一个人的知识并不只局限于知道“如何去做”,也包括看他“知道什么”。这是区分测试与考试的一个特征,因为它依靠的是“系统且彻底”的调查。传统的母语、数学、外语以及一些事实性重现领域的考试,经常在一定程度上以测试为基础,考官一眼就能看出一名学生是否正确地获取了某些知识。在拼写测试、智力测试和学校官方测试中,人们马上就能看出来某些人可以做什么,比如拼写、计算以及识别图形。这种确切的“一眼就能看出来”(seeing at a glance)意义上的测试没必要包含评级和打分。一位教师看看一个儿童写自己的名字,就可以知道他书写的情况,而此处并没有评级和打分;一位钢琴教师看看某位学生协调使用左右手的情况,就可以判断出他弹琴可以做得好或者不好。这些考查从严格意义上讲,并不是分级的问题,同样也与打分无关。

三、何为分级和评分

总结一下前文:一方面,我认为在传统的教育语言中,考试和测试或多或少可以交替使用;另一方面,考试涉及了对某些事物的“彻底调查”,而测试则与“一眼就能看出来”某物是否正确或能否达到一个标准有关。

从更为严格的意义上讲,无论是考试还是测试,都未必包含分级与评分的内容。但是,在学校教育的情景中,二者常被替换使用,基于评分的分级也更为普遍。这一点无论是在英国还是其他国家均是如此。

Hoskin 声称,他找到了评分方式首次在英国变得重要的历史时间点。他正在撰写 18 世纪晚期剑桥大学数学领域考官所面临的问题文章。当时的情况是,人们发现口试体系(oral system)管理变得越来越困难,其原因在于这一体系是建立在考官判断的基础之上,而考官的裁定经常反映的是其主观偏好,越是位高权重的人,其决定就越有分量。大约在 1792 年,时任主考人之二的 William Farish 认为,应当给每一个问题规定分数。回顾过去,这也许是朝着现实的数字化模型跨出的非常重要的一步^[2]。

分数的引入非常重要。19 世纪时,Francis Galton^[3-4]等人推动了智力测试的发展,其理念基于先天决定智力的上限是可测量且因人而异的。那种有些人天生就聪明而有些人生下来就愚笨的观念仍然深刻地影响着教师、学生以及课堂的其他组织者^[5]。

我们也都知道这种固化的能力差距是如何与社会阶层、种族、性别或地域联系起来的。在同一所学校、同一个国家中,以及整个世界范围内,学生可以通过不同的排行榜进行比较。同时,不同的社会经济组织、性别、种族背景以及地域表现也都可以进行对比和衡量,因而,搞统计的人有做不完的工作。所有这些比较可能并且经常是等级化的。A 的评分比 B 高,因而 A 在体系中就处于更高的位置或更高的层级,由此进入精英学校,待遇优厚的工作也就唾手可得。然而,这些比较是分级的产物,而非“彻底调查”意义上的考试。我们亦可以想象一个不包含分级并且排除能力固化理念的考试方法。

四、自省与考试的世俗化

到目前为止,所有的讨论都集中在考试以及测试上,这其中遗漏了对个人精神世界的透彻考察,即自省(Self-examination)。通常情况下,自省指的是通过洞察个人的思想、感受和行为,以达到对自己真正意义上的理解,从而摆脱虚荣心的驱使和其他形式的自我欺骗。值得注意的是,自省所关注的并非仅仅是某人了解或是理解了多少事情,它的涵盖面非常广泛,涉及了“这是一个怎样的人”的方方面面。从这一点上讲,它与我们所说的学术性考试大相径庭。自省流行的全盛期是一个更加宗教化的时代,即 19 至 20 世纪基督教统治时期。这一时期的欧洲和北美,人们希望时刻了解自身的精神状态。上帝作为主考人,将决定凡是上天堂还是下地狱。一般人无法知晓上帝的决定,但他们知道(或者他们认为

自己知道)的是,自己越远离罪恶就越有机会升入天堂。

有些学校现今仍在强调自省,但并不是宗教式的,它通常被叫做“自我评价”(self-assessment),稍后我再讨论。

在一个宗教化的时代中,自省意味着要接受他人审查,尤其是受神职人员的审查。牧师有时接受一般人的忏悔。自我检视与他人检视的协调是出现在 17 世纪晚期教育家夸美纽斯的论述中。“考试是对我们的虔诚进行不断的检验,可能来源于自己或是他人。我们会遇到各种诱惑以此来考验我们对上帝的虔诚,这些诱惑可能来自于人类、恶魔或者上帝。人们应当检视自己是否忠诚,是否按照上帝的旨意行事,而且我们自己被他人、朋友和敌人检视是必要的。”^[6]愚昧无知是所有可能罪恶中的一种。北欧和美洲更为虔诚的新教徒特别重视自省,他们主张人们通过《圣经》来获得关于上帝创造世界的知识,而不是像天主教声称的由宗教权威来传递知识。18 世纪英国的新教徒学院尤为重视考查这类关于上帝及其创造世界方面的知识,不仅是关于圣经的知识,还包括物质世界的知识。作为考试机构,它们要比现在的大学超前很多。由教师和校外机构主持的口试是这些学院考试制度的一个特色^[7]。因而,此时的学院考试仍有宗教印记,并被视作时刻检视自身并作为“过有道德生活”的一部分。在 19 世纪,我们看到了 Francis Galton 在 1869 年这样的论述:“我将社会和职业生活视作一个持续的考试。人们都争相获得他人的好评并取得事业上的成功”^[4]。到了 20 世纪,考试距离这种广泛的关注越来越远。在这个世俗的年代,我们只单纯地从智力层面思考考试。在 19 世纪的英国早中期,仍然将考试与道德品质联系在一起。William Gladstone(前福音派教徒,后来成为了英国财政大臣)在给女王的信中写道:“我们的大学和公立学校经验表明,公开考试同时也是有效的品格测试,先前的勤勉和克己很少与美德分开”^[8]。

那么,现在考试与道德品质的联系完全不复存在了吗?事实上,考生通常受到表扬,因为他们勤勉且有纪律性,能够妥善对待焦虑,坚持自己的目标。在英格兰,我们听到了很多 Nicky Morgan 任职教育大臣时关于品德在教育中重要作用的言论,尤其是对直面困难和克服困难等方面的强调。我想他可能没有将其与学校考试的严格性联系起来,我猜想他言外之意仍是在关注道德因素。Nicky Morgan 在 2014 年 7 月伯明翰大学的演讲中曾谈到:“确保年轻人不仅要在学业上精进,而且还要塑造品德、韧性和勇气……我们要让年轻人在毕业离校之际具备不屈不挠的品质。我们想让小学生为成就而狂欢,但同时遵守公平竞争原则,优雅地取胜,谦逊地从失败中汲取教训。”

对于品德教育的关注在我们这个国家依然强烈。你也许想探究一番当今这个被考试和测试主导了的教育系统与人格教育究竟结合到了什么程度,就如同你或许想追究现今考试取向的体系中是否还有宗教残余一样。我这里只是简单地抛出了一两个假设,而非被证明过的事实。这些观念都需要进一步的证据去证实或是证伪。

到目前为止,希望我已经说明了考试比我们初看之下要复杂得多,这样一个事实。在历史上很长一段时间内,传统考试都不是单纯地对个体知识进行评估,还有促进其品德发展的功能。如果我们想对考生所学科目的理解程度作一个彻底的调查,那么仅仅进行评价知识的考试是不够的,是存在缺陷的。从一个严格的意义上讲,我们所侧重的只是考试中一些无关紧要的东西,例如分级和评分。这就引出了另一个问题,即我们是否(或在何种程度上)应该继续坚持传统的考试。

五、传统考试带来的问题及思考

(一)传统考试会导致个人苦恼

我已经谈到了一些考试可能会带来的焦虑。在一些极端的例子中,考试焦虑将会是灾难性的。在中国香港、印度以及英国,我们最近了解到一些年轻人因为担忧考试而产生自杀的念头,这种想法会导致自残甚至是真正的自杀。我们都认同一个原则,即不应伤害他人。我们保护年轻人,使他们免于遭受虐待和肢体残害,我们反对教师体罚,反对同学间的欺凌以及课堂内的冷暴力。然而,我们却不觉得考试焦虑所带来的伤害可以比肩虐待、体罚、欺凌以及冷暴力,这是为什么呢?

(二)传统考试限制了学校课程

考试的基石是学校科目,比如数学、地理以及法语。为了以后能够在考试中获得成功,多数的中学乃至小学的课程都围绕具体的科目展开。我们早已经习惯了考试主导的学校教育,因而看不到这样规划课程有什么匪夷所思之处。一个恰当的解释是学校的教育目的及其衍生出来的具体目标,成为考试主导学校教育的理由^[9]。唯有如此,我们才能知晓什么样的机制能够达成目标。有些时候,跨学科的工作(interdisciplinary work)比单个学科更加值得提倡。它可以让学校从考试主导的境遇中解脱出来,也可以带来更有想象力、更富吸引力的学习方式。

(三)传统考试带来的其他不良状况

考试还带来了许多其他不良状况。譬如:教师的应试化教学方式,学生的应试化学习方式,考试中作弊现象的大量出现,等等。另外,许多学校在考试方面投入了大量的时间、精力和经费,但所得到的利益(benefit)还没有大学和未来的雇主(employers)多。此外,考试造成人们对自己和对他人不适当的感受和看法。不管是儿童还是成人都认为自己天生就处于某个能力水平,或是聪明或是愚笨,其实这仅与优越感或低自尊心有关,并没有足够确切的证据说明能力是一成不变的。

(四)传统考试引发的思考

我们不禁会发出疑问:为什么很多国家看到了传统考试的诸多缺陷,却仍抱残守缺?据称传统学校考试的一个主要功能就在于维持某种社会结构,即那种精英阶层能够代代相传的等级社会。

结合英国或是其他国的考试政策历史,这点在多大程度上能够得到印证呢?就拿英国来说,考试适用于 1854 年的英国文官制度。随着工业革命的进行,日益壮大的中产阶级发起了一项旷日持久的运动,其目的就是要在国家治理中扮演更重要的角色。地主阶级在当时占据主导,借助一种裙带系统,他们把自己的人置于领导位置,而新兴的中产阶级则认为公开考试是一个更为公正的系统。这一体系在上世纪 60 年代之前一直存在。在一个多世纪的时间里,只有少数私立学校的学生可以通过学校考试进入大学深造(从 1902 年起,公立文法学校的学生也可以),并且这些学生都来自于上层社会的家庭。其余大部分其他学校的学生均被排除在考试之外,而他们中的大多数来自于较底层社会的家庭。1988 年引入了国家课程,因而自那时起我们开始适应一种更加平等的教育体系。事实上,每个人都有望获得普通中等教育证书,并且如果他们还有能力的话可以去考 A-level 并进入大学。有些人认为,这一体系只是看起来平等而已,但实际上更高社会层级的人更有能力操纵体系,让他们的子女比其他人获得更好的考试分数。他们会让子女接受私立学校的教育,实施小班化教学,硬件设备丰富;或者也会让其子女进入教育质量优质的公立学校,并搬到这些学校的学区,雇佣私人家庭教师。

因此,这里有三种假设供仅参考:一是英格兰 1988 年以来形成的体系的确是平等的,它为每个人都提供了在重要考试中取得成功的均等机会;二是自 1988 年起,政府想创立一种看起来是平等的体系,但实际上却偏向富裕阶层;三是富人确实从 1988 年后的体系中获利,然而这并不是精英阶层蓄谋已久的政治行为的结果,它受望子成龙、望女成凤的小康家庭观念影响更大。那么,哪一种假设更有事实依据呢?

六、一种新的评价方案出现

Alison Peacock 出版了一部新书,名为 *Assessment for Learning without Limits* (2016)。她是伦敦附近一所小学的校长,同时也是一个名为“学习无限制”(Learning without Limits)研究项目的领军人物。这一项目否定那种能力是有固定局限的说法,而这个说法的支持者是 Galton 和 Burt 等人。这部专著基于她在其学校中的亲身评估实践,其中的基本理念之一就是不为学生打分,不鼓励学生将自身的表现与他人进行比较。因为这些对比强化了一个观念,即一些人天生就比其他人的能力强,现在如此,将来亦是如此。因此,她的评价方式与我们所熟知的评价方式是截然不同的。

新的评价方案对人的评价变得更加复杂和微妙,并且具有持续性和发展性,它的目标在于让学生不断地学习,挑战自我并达到更高的成就。由于学生之间并不相互竞争、攀比分数,所以重点就放在了合作性学习和相互帮助上面。这种评价并不是由权威人物掌控(比如教师),因其本身是一种合作性活动,所以参与的人员不仅有教师,而且还包括学生本人及其家长,通过会议及在线交流,所有人都可以探讨每位学生的表现,消除误解,增强学生的信心,从而帮助学生达到更高的成就。学生的自我评价在此处也很重要,他们的成就通过文字、图片和视频的方式分享出来,构成了对其成就的一种持续性记录。

在本文开头我便提出,从严格的意义上来讲,考试应是一种“彻底的调查”。在教育领域,它是对人们所知晓、所理解以及所作所为的一种“彻底的调查”。在我看来,Alison Peacock 所描绘的考核形式与此目标更加契合。相较于传统的考试而言,它展示了一种对学生的行为和他们所掌握的知识更为深入的调查方法,因而是一种严格意义上的考试。

参考文献：

[1] DAVIS A. Educational Assessment: A Critique of Current Policy[J]. Impact, 1999, 9(1): 49.
[2] KEITH H. The Examination, Disciplinary Power and Rational Schooling[J]. History of Education, 1979, 8(2):21-135.
[3] HOSKIN K. Foucault Under Examination: The Crypto-Educationalist Unmasked[M]//Ball S. Foucault and Education: Disciplines and Knowledge. London:Routledge,1990:29-53.
[4] Deary I J. Intelligence, Destiny and Education: The Ideological Roots of Intelligence Testing[J]. Intelligence, 2006, 34(6):621-622.
[5] PEACOCK A. Assessment for Learning without Limits[M]. London: London Open University Press,2016:5.
[6] COMENIUS J A. The Great Didactic[M]. London: Adam and Charles Black, 1638:82.
[7] WHITE J. The Invention of Secondary Curriculum. Secondary Education in a Changing World[M]. New York: Palgrave Macmillan,2011:72-74.
[8] MAGNUS P. Gladstone:Biography[M]. London: John Murray,1954:117-118.
[9] REISS M, WHITE J. An Aims-based Curriculum: The Significance of Human Flourishing for Schools[M]. London: IOE Press, 2013.

Reflections on the Notions of Examining, Testing, Marking and Grading

Written by John White¹, Translate by Zhao xiantong²

(1.UCL Institute of Education, London England WC1H0AL;

2. Department of education, southwest university,Changqing 400715, China)

Abstract: The paper begins by investigating the concepts of examining, testing, grading and marking and relations between them. It shows that what we call ‘examinations’ in the educational world can be very different from an examination as systematic and thorough investigations and how the notion of marking became central to how we think of examinations. The paper then looks at self-examination, its rootedness in religious concerns and links that have been made between examination success and possessing moral virtues. A section on a range of problems with conventional examinations is followed by a final section that looks at a positive alternative. This brings examining in school closer to the notion of systematic and thorough investigation.

Key words: examining; testing; grading and marking; self-examination; elite schooling; formative assessment

责任编辑 邱香华