计分方式及选考群体对学生选科成绩的影响

崔 伟 殷 乐 李晓庆 吕 啸

【摘 要】本研究利用学生考试数据，模拟了等级赋分、标准分和调整标准分三种计分方式及高水平、低水平两种选考群体对学生选科成绩的影响。研究发现：标准分和等级赋分的计分方式会导致学生在不同选考群体所得成绩差异巨大，引发学生“躲避好学生”的倾向，不宜作为学生选考成绩计算方法；调整标准分计分方式也会导致学生成绩随选考群体发生变化，但影响趋势与标准分、等级赋分相反，可能引发学生“追随好学生”的倾向。对此，可针对学生不同选考阶段采用不同计分策略：学科探索阶段采用标准分；升学考试中的选考科目采用调整标准分；录取阶段多学科分别计分，由高校独立赋权计算总分取代各科直接简单相加计算总分的录取模式。

【关键词】考试计分；标准分；等级赋分；调整标准分；模拟分析
【中图分类号】G632.474 【文献标识码】A 【DOI编码】10.16518/j.cnki.emae.2018.03.010

一、问题提出

2014年9月，《国务院关于深化考试招生制度改革的实施意见》颁布，标志着新一轮考试招生制度及教育评价改革全面启动。此次改革的突出特点是取消了文理分科，学生可以自主选择选考科目参加升学考试，考试总成绩由统考科目成绩和选考科目成绩组成。面对学生选考科目的差异，如何计算选考学科的成绩，实现不同选考学科成绩具有可比性、不同选考组合成绩具有可比性，是非常重要的课题。同时，考试分数在录取时所发挥的重要作用，也使得学生对成绩变化非常敏感。因计分方式带来的成绩波动对学生的选科具有巨大影响，可能使学生的“自主选科”变成“计分选科”。

回顾我国高考计分历史，已出现了三种计分方式：原始分（也称卷面分）、标准分和等级赋分。原始分是试卷上所有小题得分相加之和，因而具有简洁、明了、方便计算等优点。但原始分的单位和参照点会因测验难度的不同而不同，欠缺稳定性，也不具备可加性。[1]为克服原始分的缺点，1985年，国家开始在广东省试点高考标准分，并逐渐推广至20个省市。标准分是依据统计与测量的原理，将原始分转化成具有相同单位和共同参照点的分数，从而使不同科目成绩具有可比性、可加性。[2]但标准分的计算过程较复杂，不易被一般民众理解。2001年后除了海南省还在采用标准分，各地高考纷纷回归原始分。这种回归，有学者提出是因为“3+X”高考政策与原标准分计算方式不相容。[3]具体来看，选考背景下，忽视选考群体能力差异，直接用各选考群体的原始分来计算标准分是不合理的。等级赋分是本次高考改革第一批试点的浙江省首先提出的。选考科目成绩以当次学考合格成绩为赋分前提，按选考人数比例划分21个等级，起始赋分为40分，满分100分，相邻两个等级相差3分。[4]等级赋分的思想与标准分一致：统一各学科成绩的单位和参照点，根据学生成绩在群体中的位次重新计算分数。文

崔 伟/北京师范大学未来教育高精尖创新中心研究员，博士，主要研究方向为学习分析，心理、教育测量，学业发展规划。（北京 100875）
殷 乐/北京师范大学未来教育高精尖创新中心研究员，博士，主要研究方向为心理、教育测量。
李晓庆/北京师范大学未来教育高精尖创新中心学科教育实验室常务副主任，主要研究方向为大数据助力区域教育质量改进、信息技术与课程深度整合研究。
吕 啸/供职于北京师范大学出版社，主要研究方向为高考数据分析。

东茅等人通过模拟分析，确认等级赋分具有可以接受的区分度[5]，但分析过程并未考虑选考群体差异可能带来的影响。从试点结果来看，浙江省也出现了物理选考人数锐减的现象，与当年广东省“3+X”情况一致，学生的“田忌赛马”选考策略非常突出。[6]对此，有学者提出在选考背景下，应采用调整标准分计算学生成绩[7]，认为该计分方式对“强—强”（能力强的学生在强能力组）、“强—弱”“弱—强”“弱—弱”都比较公平。[6]

当前，关于选考科目成绩究竟如何计分，各地方案不同。第一批试点地浙江省、上海市为等级赋分，第二批试点地海南省采用标准分。[8]而学者们的讨论多以理论分析为主，真实数据的模拟研究极少，难以直观显示选考背景下，不同计分方式及选考群体对学生选科成绩的影响。因此，本研究利用学生考试数据，模拟标准分、等级赋分、调整标准分等当前被讨论较多的计分方式，探讨选考群体和计分方式对学生最终选科成绩的影响，以期为选考科目的计分方案提供有价值的参考。

二、数据来源和模拟分析方法

（一）数据来源

本研究采用北京市某区初二学生的期末考试数据，利用 Excel 2016 和 SPSS 24.0 软件进行分析。考试科目包括语文、数学、英语、物理、生物、地理、历史、思想品德（简称思品），其中英语满分 70 分，其他科目满分均为 100 分。试题由区教研员统一命制，采用网上匿名判卷、集体流水阅卷方式，对主观性较高的题目则采用复评机制。考生总人数为 4340 人，剔除缺考一门及多门学生，实际用于统计分析的考生为 4253 人，其中男生 2261 人，女生 1922 人，性别未知者 70 人。各学科原始分的描述性统计见表 1。

表 1　八个学科原始分情况统计

	语文	数学	英语	物理	生物	地理	历史	思品
平均值	78.9	68.6	38.9	73.9	62.6	68.1	66.8	70.0
标准差	11.9	21.7	15.7	15.7	16.1	17.1	16.6	10.3

（二）模拟分析过程和方法

1.计算各科标准分成绩：①利用 SPSS 分别将 4253 名学生的学科成绩分别进行正态化转化并计算学生的 Z 分数；②利用公式“标准分=500+100×Z”，计算学生每个科目的标准分。

2.形成不同选考群体：因语文、数学、英语为统考科目，可以获得所有学生的成绩数据，故以此 3 科的成绩为依据确定不同选考群体。①将 3 科的标准分成绩相加后，经正态化转化，计算 3 科总成绩的标准分；②按 3 科总成绩的标准分，由高到低对学生排序；③随机从前 3000 名学生中选择 2000 名组成“高水平”选考群体（以 A 表示）；④随机从后 3000 名学生中选择 2000 名组成“低水平”选考群体（以 B 表示）。

3.选择“被观察”学生模拟选科：①从样本总体中随机选取 30 名学生，组成“被观察”学生（以 C 表示）；②将“被观察”学生分别加入 A 组（即学生加入“高水平”选考群体情况）和 B 组（即学生加入“低水平”选考群体情况），模拟具体学生的选科。

4.计算选科得分：以物理学科为学生选考学科，分别按照浙江、上海两地的等级赋分规则，标准分、调整标准分的计算方法，计算学生的选考科目得分结果。

等级赋分：①浙江省赋分以及格为赋分起点，因此先删除物理不及格学生，再将选科学生按比例划分为 21 个等级进行物理赋分，起始赋分为 40 分，满分 100 分，相邻两个等级相差 3 分[4]；②上海市赋分则是根据规定将选科学生按比例划分为 11 个等级，起始赋分为 40 分，满分 70 分，相邻两个等级相差 3 分。[9]

标准分：①总体标准分是先将所有考生（4253 人）的物理原始分进行正态转化，再计算其 Z 分数，最后按标准分计算公式计算物理的总体标准分；②群体标准分是先将选考群体所有学生（2030 人）的物理原始分进行正态转化，再计算学生的群体 Z 分数，最后按标准分计算公式计算学生群体标准分。

调整标准分的计算步骤（以 A+C 组成的高水平选科群体为例）[7]：①将 A 组和 C 组（2030

人）的语文、数学、英语3科标准分相加，再将成绩转化为正态分布下的标准分，作为该群体统考基础分；②计算统考基础分的均值和标准差，作为该群体选考科目物理的均值和标准差；③将该群体学生的物理原始分进行正态转化，并计算学生的Z分数；④调整标准分=统考均值+统考标准差×Z。可以看到，调整标准分的一个重要假设是学生统考科目成绩与选考科目成绩具有较高正相关。本次分析数据中统考科目标准分与其他各科目标准分的相关均较高（见表2），满足上面的假设。

5.对比两个选考群体在不同计分方式下的选科成绩。

6.对比“被观察”学生在不同选考群体、不同计分方式下的选科成绩。

三、模拟结果

（一）不同计分方式下，学生选科成绩描述性统计

表3和表4分别为根据浙江省和上海市等级赋分规则，两个选考群体在不同等级的划界分。其中，浙江省赋分，按规则先删除物理原始分不及格的学生，A+C组共1932名学生参与赋分，B+C组共1505名学生参与赋分。可以看到，按浙江省和上海市的赋分规则，在各个等级的划界分，A组均高于B组。根据浙江省赋分规则，A组划界分比B组划界分平均高出5.5分，两组分差较大的集中在中间等级（8~16级）。根据上海市赋分规则，A组划界分比B组划界分平均高出11分，等级越低两组划界分差值越大。之所以会出现此结果，可能是因为浙江省赋分规则中将不及格的学生删除，上海市则未删除该部分学生。

表5为两个群体学生在不同计分方式下的选科成绩情况。其中总体标准分是指以4253名学生的物理成绩为基准计算所得标准分，群体标准分是指以各自的选考群体中2030名学生的物理成绩为基准计算所得标准分。可以看到原始分、总体标准分和调整标准分下的两组成绩不一致，A+C组成绩高于B+C组成绩；群体标准分、浙江赋分、上海赋分下两组成绩的平均值和标准差都非常接近。

表6为30名“被观察”学生在不同群体、不同计分方式下的选科成绩情况。可以看到，群体标准分、浙江赋分、上海赋分时，C组学生在A+C

表2　统考科目（语、数、外）总标准分与选考科目标准分的相关

	物理标准分	生物标准分	地理标准分	历史标准分	思品标准分
皮尔逊相关性	0.790***	0.710***	0.760***	0.767***	0.734***
p	0.000	0.000	0.000	0.000	0.000
n	4253	4253	4253	4253	4253

注：*** 表示在0.001水平（双尾），相关显著。

表3　根据浙江省赋分规则的两组划界分情况

等级	1	2	3	4	5	6	7	8	9	10	11
赋分	100	97	94	91	88	85	82	79	76	73	70
A+C组划界分	99	97	96	95	93	92	90	89	87	85	83.5
B+C组划界分	97	94	92	90	88	86	84	81	79	77	75
等级	12	13	14	15	16	17	18	19	20	21	
赋分	67	64	61	58	55	52	49	46	43	40	
A+C组划界分	81	79	77	74	72	69	66	64	62	61	
B+C组划界分	73	71	69	67	65	64	62	61	60	60	

表4　根据上海市赋分规则的两组划界分情况

等级	1	2	3	4	5	6	7	8	9	10	11
赋分	70	67	64	61	58	55	52	49	46	43	40
A+C组划界分	99	95	91	89	87	84	81	78	74	69	60
B+C组划界分	97	90	84	80	76	72	68	64	59	52	40

表 5　两组学生在不同计分方式下的选科成绩

	原始分	总体标准分	群体标准分	浙江赋分	上海赋分	调整标准分
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
A+C 组	80.62 (10.98)	540.80 (82.33)	499.93 (99.66)	70.52 (13.56)	54.58 (8.70)	549.38 (72.19)
B+C 组	68.30 (15.31)	461.79 (85.88)	499.97 (99.82)	70.67 (13.68)	54.70 (8.61)	452.02 (71.90)

表 6　“被观察”学生在不同群体、不同计分方式下的选科成绩情况

	原始分	总体标准分	群体标准分	浙江赋分	上海赋分	调整标准分
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
C_A	66.20	513.75	464.33 (153.77)	72.75 (15.11)	53.40 (10.74)	523.60 (111.39)
C_B	(21.68)	(124.15)	561.48 (143.86)	81.75 (14.77)	59.20 (10.76)	496.33 (103.63)

注：浙江赋分，30 名学生中有 6 名成绩低于 60 分，根据赋分规则删除，剩余 24 名学生参与统计。

组所得成绩（C_A）均低于在 B+C 组所得成绩（C_B）；调整标准分时，C 组学生在 A+C 组所得成绩（C_A）高于在 B+C 组所得成绩（C_B）。

（二）两个选考群体在不同计分方式下的选科成绩比较

表 7 为采用独立样本 t 检验对比两个群体在不同计分方式下的选科成绩比较结果。我们发现，两个群体的原始分、总体标准分和调整标准分，A+C 组平均分均显著高于 B+C 组，$t_{原始分}=-29.457$，$df=3679.6$，$p<0.001$；$t_{总体标准分}=-29.921$，$df=4058$，$p<0.001$；$t_{调整标准分}=-43.052$，$df=4058$，$p<0.001$；两个群体的浙江赋分、上海赋分及群体标准分，A+C 组与 B+C 组无显著性差异，$t_{浙江赋分}=0.301$，$df=3435$，$p=0.763$；$t_{上海赋分}=0.473$，$df=4058$，$p=0.636$；$t_{群体标准分}=0.016$，$df=4058$，$p=0.988$。也就是说，从原始分和总体标准分来看，A+C、B+C 两组确实是两个不同质的群体，A+C 组在物理学科的学业成绩明显好于 B+C 组。调整标准分能够体现出两组间的差异，但群体标准分、浙江省等级赋分、上海市等级赋分均未能体现出两组间的差异。

（三）“被观察”学生在不同群体、不同计分方式下的选科成绩比较

30 名 C 组学生在两个选考群体（A+C 组、B+C 组）、四种计分方式下的选科成绩对比见表 8。我们采用配对样本 t 检验分别比较四种计分方式下，C 组学生在 A+C 组所得选科成绩（C_A）与在 B+C 组所得选科成绩（C_B）是否存在显著性差异，得到如下结果。①群体标准分计分：成绩 C_A 显著低于成绩 C_B，$t=-43.382$，$df=29$，$p<0.001$，成绩差值为 97.15 分，接近标准分计算的一个标准差（100）。根据正态概率分布，平均值附近一个标准差的人数概率分布约为 38%。②等级赋分：依据浙江省赋分规则，成绩 C_A 显著低于成绩 C_B，$t=-18.838$，$df=23$，$p<0.001$；依据上海市赋分规则，成绩 C_A 显著低于成绩 C_B，$t=-7.918$，$df=29$，$p<0.001$。③调整标准分计分：成绩 C_A 显著高于成绩 C_B，$t=15.963$，$df=29$，$p<0.001$，成绩差值为 27.27 分，较群体标准分差值缩小。

也就是说，标准分和等级赋分时，学生在高水平选科群体中所得选科成绩要低于在低水平

表 7　两组学生在不同计分方式下的选科成绩比较

	人数	差值	差值标准差	*t*	*df*	*p*
原始分	4060	12.32	0.418	−29.457	3679.60	0.000
总体标准分	4060	79.01	2.64	−29.921	4058	0.000
群体标准分	4060	0.05	3.13	0.016	4058	0.988
浙江赋分	3437	0.14	0.47	0.301	3435	0.763
上海赋分	4060	0.13	0.272	0.473	4058	0.636
调整标准分	4060	97.36	2.26	−43.052	4058	0.000

注：根据浙江赋分规则删除物理不及格个案，A 组删除 98 人，B 组删除 525 人，共剩余 3437 人参与分析。

表 8　“被观察”学生在两个选考群体、四种计分方式下的选科成绩比较

C_A-C_B	人数	差值	差值标准差	t	df	p
群体标准分	30	–97.15	2.239	–43.382	29	0.000
浙江赋分	24	–9.00	2.341	–18.838	23	0.000
上海赋分	30	–5.80	4.012	–7.918	29	0.000
调整标准分	30	27.27	9.357	15.963	29	0.000

注：浙江赋分，30 名学生中有 6 名物理成绩低于 60 分，根据赋分规则删除，剩余 24 名学生参与统计。

选科群体中所得选科成绩，即群体中“好学生”越多，学生成绩越低；“好学生”越少，学生成绩越高。调整标准分计分下，学生在高水平选科群体中所得选科成绩要高于在低水平选科群体中所得选科成绩，即群体中“好学生”越多，学生所得成绩越高；“好学生”越少，学生所得成绩越低。

四、结论与思考

本研究模拟了不同计分方式、不同选考群体对学生选科成绩的影响，发现计分方式和选考群体均会对学生选科成绩产生重要影响，从而影响学生的选科策略。

具体来说，群体标准分和浙江、上海两地的等级赋分方式均忽视了选考群体间的差异，高水平选科群体与低水平选科群体被看做是相同的群体，两组间的差异被计分方式消除。同时，选考群体对学生个人的选科成绩产生了重要影响，在高水平群体所得成绩远低于在低水平群体所得成绩。假设物理选考群体的“好学生”集中，而化学选考群体中“好学生”较少，某个学生的物理、化学均处于总体的平均水平，最终其选考物理与选考化学成绩将差异巨大，这必然导致其放弃“好学生”集中的物理而选择化学，从而出现“田忌赛马”选科策略。实际选考中，“好学生”的确更倾向选择物理。可以说，群体标准分和等级赋分的确是广东、浙江两地物理选考人数急剧回落的重要原因之一。[6][7]调整标准分能够体现选考群体间的差异，高水平选科群体的整体选科成绩仍高于低水平选科群体。同时，选考群体对学生调整标准分成绩也有影响，但影响较群体标准分小，且影响方向与标准分计分相反：群体中“好学生”越多，学生选科成绩越高；“好学生”越少，选科成绩越低。当前，学生取得最高成绩的动机非常强烈，三种计分方式对学生选科策略的影响是：标准分计分和等级赋分时，学生会倾向于“躲避好学生”；调整标准分计分时，学生会倾向于“追随好学生”。

本文的分析表明，在选考阶段采用标准分计分和等级赋分均有失公平，且存在学生“躲避好学生”的倾向。调整标准分较好地维持了群体间差异，但学生成绩仍然受到选考群体的影响，存在学生“追随好学生”的可能，或许可以发挥好学生的引导作用。虽然调整标准分并未很好地解决选考计分的等值问题（C 组学生在两个群体的调整标准分仍存在显著差异）[10]，但与群体标准分、等级赋分相比更具合理性。

根据模拟分析结果，结合当前标准化考试和等值性测验仍然匮乏的现实，为了实现选考政策的价值取向，帮助学生发现自我优势、学有所长，教育管理部门应针对不同选考阶段选用不同计分策略。

首先，在学生学科优势探索、学科兴趣形成阶段，宜采用标准分形式报告学生成绩。此阶段，所有学生参与所有学科的学习和考试，群体差异较小。标准分计分消除了学科间难度和内容上的差异，能够帮助学生较准确地识别自身学科优势。

其次，在升学考试阶段，宜采用调整标准分。此阶段，学生选考学科不同，既有学科差异，又有群体差异。调整标准分能平衡学科差异，同时能维持群体差异，降低选考群体对学生成绩的影响，使得不同学科、不同群体学生成绩具有一定可比性。

最后，在录取阶段，改革总分录取计分方式，采用多学科分别计分。学生选考科目不同，所在群体不同，各科成绩简单相加显然科学性不足，

也抹杀了学生选考学科特征。教育行政部门可将学生统考学科成绩、选考单科成绩及所在选考群体特征数据同时提供给高校，由高校根据自身专业特点对统考、选考学科赋予不同权重，选择合适的人才。

需要说明的是，以上分析及结论仅基于本文数据及假设学生采用“成绩最高”选考策略。现实中，选考决策是受众多因素影响的复杂决策，除了成绩因素，个人兴趣志向、学校、家庭都会影响学生的最终选择。因此，本文结论是否具有普遍意义仍需进一步研究。此外，虽然本文所用数据满足统考学科成绩与其他学科高相关的假设，但若某次考试不符合该假设，调整标准分将导致更大偏差。

参考文献：

[1] 杨志明.高考原始分合成：问题与改进思路[J].教育测量与评价，2015(10)：61-64.

[2] 章建石.一项公平与效率兼备的高考改革为什么难以为继?——标准分制度的变迁及其折射的治理困境[J].北京师范大学学报(社会科学版)，2016(1)：31-41.

[3] 温忠麟.高考改革：政策公平性与技术相容性[J].全球教育展望，2014，43(2)：3-14.

[4] 浙江省教育厅.浙江省教育厅关于印发《浙江省普通高中学业水平考试实施办法和浙江省普通高校招生选考科目考试实施办法》的通知[EB/OL].(2014-11-07)[2017-12-13].http://www.zjedu.gov.cn/news/27105.html.

[5] 文东茅，鲍旭明，傅攸.等级赋分对高考区分度的影响——对浙江“九校联考”数据的模拟分析[J].中国高教研究，2015(6)：17-21.

[6] 柯政.“选考”制度下的“田忌赛马”：原因与对策[J].教育发展研究，2016(18)：32-38.

[7] 温忠麟，罗冠中.高考“3+X”分数转换和总分合成方法[J].考试研究，2006(3)：43-45.

[8] 海南省教育厅.海南省教育厅关于印发《海南省普通高中学业水平考试实施办法》和《海南省完善普通高中学生综合素质评价实施办法》的通知[EB/OL].(2016-03-20)[2017-12-13].http://edu.hainan.gov.cn/news-article-18352-1.html.

[9] 上海市人民政府.上海市人民政府关于印发《上海市深化高等学校考试招生综合改革实施方案》的通知[EB/OL].(2014-09-18)[2017-12-13].http://old.moe.gov.cn//publicfiles/business/htmlfiles/moe/s8367/201409/175288.html.

[10] 杨志明.一年多考背景下分数等值的意义和方法[J].教育测量与评价，2015(12)：58-61.

The Influence of the Scoring Method and the Selected Groups Acting on Students' Scores of Choosing Subjects

Cui Wei，Yin Le，Li Xiaoqing，Lü Xiao

Abstract: This study used the data of students' examination to simulate three scoring methods: graded scoring，standard score，adjusted standard score and analyzed the influence of high level and low one among the selected groups acting on the students' scores of choosing subjects. The results show that the standard score and graded scoring are not suitable to be used because they will lead to huge difference in students' scores in different selected groups，which causes the tendency of avoiding good students for some students. The adjusted standard score can also cause students' scores changing with the selected groups，but the influence trend is opposite to above one，which brings about the tendency of following good students for some students. Thus，this paper suggests that different scoring strategies should be adopted at different stages：the standard score could be used at the exploratory stage；the adjusted standard score could be used in entrance examinations；multi-discipline score instead of a single total score should be applied at admission stage from the mode of total score adding up by each subject simply to that of independent weighting by colleges and universities.

Keywords: test scoring，standard score，graded scoring，adjusted standard score，simulation analysis

责任编辑/王彩霞