

项目难度与项目极大区分度之间的关系

席仲恩

(作者简介: 席仲恩, 副教授, 上海外国语大学博士生, 绍兴文理学院语言测试与评价研究所所长, 上海, 200083)

摘要: 本文在项目就是待测量特质的量具、受考在项目上的作业成绩为正态分布这两个简单假定的条件下, 通过简单的数学推导, 确定出了经典测试理论模型中的项目难度(用易度指数表示)和极大区分度(用高低分组通过率或得分率差表示)之间的两个函数关系式, 望这两个关系式能为题库建设和试卷开发提供一个既明确又简单可行的参考数学模型。

关键词: 心理测验; 教育测量; 项目分析; 项目难度; 项目极大区分度

一、引言

项目分析(item analysis)是题库建设(item banking)的至关重要的一个环节, 是开发高质量试卷的基础性工作。试卷由一道一道的项目组成, 如果没有合格的项目, 就不可能有合格的试卷; 没有高质量的项目, 就无法构成高质量的试卷。对于一个项目, 既可以做定性分析, 也可以做定量分析。本文关心的仅是项目的定量分析。

对于项目进行定量分析时, 在经典理论框架中, 通常求出项目的两个指标: 难度指标和区分度指标。难度指标传统上用项目的易度指数(facility index)——项目的平均答对率(对于选择题)或平均得分率表示^①。区分度指标尽管有多种表示方法, 本文选取了用高低分组难度差刻划的区分度, 一方面这种区分度直观, 其含义又和项目反应理论框架下的区分度一致(请参见席仲恩, 2001), 而且经过爱贝尔(Ebel)的研究提出了具体的项目评价指标(参见 Ebel & Frisbie, 1986: P234), 此外它计算又十分简便, 便于应用。

在我国现行的心理与教育测量著作(例如: 戴忠恒, 1987; 王孝玲, 1989; 谢小庆, 1988; 余嘉元, 1987; 于信凤, 1987)和语言测试论著(例如: 高兰生、陈辉岳, 1996; 桂诗春, 1986; 李筱菊, 1997; 刘润清、韩宝成, 2000; 舒运祥, 1999; 徐强, 1992; 杨钟琳 1992)、国外的同类著作(例如: Anastasi, 1976; Salvia & Ysseldyke, 1996; Alderson, Clapham & Walt, 1995; Bachman, 1990; Bachman and Palmer, 1996; Baker, 1989; Harris, 1969; Heaton, 1988; Lado, 1961; Madsen, 1983; Spotsky, 1995)中, 都未见专门详细定量论述项目极大区分度和项目难度之间的关系的。使题库建设和试卷开发者无所适从, 给工作带来一些不便。导致像中国全国硕士研究生英语入学考试这类重大筛选性考试中难度、区分度失调, 或者只顾难度而忽略区分度等问题, 使试卷中的项目区分度普遍不高, 甚至出现负值这种奇特现象, 严重威胁了考试的信度和效度, 使考试的功效打了折扣。本文在项目就是待测量特质的量具、受考在项目上的作业成绩为正态分布这两个简单假定下, 通过简单的数学推导, 确定出了经典测试理论模型中的项目难度和极大区分度之间的函数关系, 望能为题库建设和试卷开发提

供一个既明确又简单可行的参考模型。

二、关系式的推导

本文关心的是数学模型的建立，并不关心实施中的问题。但是，这并不意味着我们认为实施中的问题就不重要，相反，我们认为实施中的问题一样重要。我们这样做，是由我们的研究目标决定的。我们的目标是建立一个理想状况下的数学模型，为实践提供一工作参考和努力方向，使实践者明确项目和试卷开发的潜力和限度，更好地发挥和挖掘考试的功效。为此，我们假定已经获得了有关项目的可靠的实际测量数据，而且这些数据的采集程序可靠、准确、合法。

设 DI 为项目区分度， U 为高分组（一般指总分最高的 27% 或 30% 受考）的答对率或平均得分率， L 为低分组（一般指总分最低的 27% 或 30% 受考）的答对率或平均得分率，则高低分组难度差区分度可定义为：

$$DI=U-L \quad (1)$$

设 FI 为表示难度的易度指数， M 为中间组的答对率或平均得分， U_{MAX} 为高分组的极大答对率或极大平均得分率， L_{MIN} 为低分组的最小答对率或最小平均得分率，假定受考在给定项目上的作业成绩为正态分布，则显而易见，项目的易度指数可定义为：

$$FI=M=\frac{U+L}{2}=\frac{U_{MAX}+L_{MIN}}{2} \quad (2)$$

同时，我们又设 DI_{MAX} 为项目的极大区分度，则极大区分度显然可以定义为：

$$DI_{MAX}=U_{MAX}-L_{MIN} \quad (3)$$

再由 (2) 中析出

$$FI=\frac{U_{MAX}+L_{MIN}}{2} \quad (4)$$

联立解由 (3)、(4) 组成的方程组得：

$$DI_{MAX}=2U_{MAX}-2FI \quad (5)$$

$$DI_{MAX}=2FI-2L_{MIN} \quad (6)$$

表达式 (5)、(6) 就是我们要求的项目极大区分度和用易度指数表示的项目难度之间的关系式。两个式子沿不同的方向（一个从上而下，一个从下而上）刻划了易度与极大区分度之间的函数关系。在项目的易度指数值大于或者等于诸极大区分度中最大的那个极大区分度（以下称做“最大极大区分度”）所对应的易度指数值时，可根据公式 (5) 计算出对应于不同易度指数值的极大区分度；在项目的易度指数值小于或者等于最大极大区分度所对应的易度指数值时，可根据公式 (6) 计算出对应于不同易度指数值的极大区分度（见表 1 和表 2）。显而易见，最大极大区分度既可以根据公式 (5) 求得，也可以根据公式 (6) 求得。

由于在项目易度指数达到最大极大区分度所对应的易度指数时， $U_{MAX}=1$ ，以及 $L_{MIN}=1/A$ （ A 为备选项的个数），则 (5)、(6) 简化为

$$DI_{MAX}=2-2FI \quad (7)$$

$$DI_{MAX} = 2FI - 2/A \quad (8)$$

(7)、(8)两式就是我们要建立的项目难度与项目极大区分度之间的函数关系。

三、关系式的应用

关于公式(7)、(8)的应用可分有猜测因素存在和无猜测因素存在两种情况。两者的区别主要在A的取值上。如果没有猜测因素存在,就相当于 $A=\infty$;如果有猜测因素存在,备选项个数是几,A就取几。显而易见,有猜测因素存在时,低分组的最小答对率就是假定这组受考对该项目一无所知,仅凭随机猜测就能答对该项目的概率,有关扣除这个因子的理据,由于与本文关系不大,恕不赘述(感兴趣的读者,请参看席仲恩,2000b)。以下,我们先计算有猜测因素存在时的极大区分度,然后再计算没有猜测因素存在时的极大区分度。

(一)有猜测因素存在时易度与极大区分度的关系

我们计算 $A=2$ (即二选一), $A=3$ (即三选一), $A=4$ (即四选一),和 $A=5$ (即五选一)时,项目易度与对应的项目极大区分度。因为,这几种选型是最常见的多选题项目。我们给出 $1\sim 0.20$ ^②之间不同易度指数值,为了方便起见,变化幅度基本取0.05,到了最大极大区分度对应的易度指数时,我们给出它的确切值。求出的相应极大区分度见表1。

表1 考虑猜测因素时难度与极大区分度之间的关系

易度指数值 FI	极大区分度 $A=2DI_{MAX}$	极大区分度 $A=3DI_{MAX}$	极大区分度 $A=4DI_{MAX}$	极大区分度 $A=5DI_{MAX}$
1.00	0.00	0.00	0.00	0.00
0.95	0.10	0.10	0.10	0.10
0.90	0.20	0.20	0.20	0.20
0.85	0.30	0.30	0.30	0.30
0.80	0.40	0.40	0.40	0.40
0.75	0.50	0.50	0.50	0.50
0.70	0.40	0.60	0.60	0.60
0.667	0.334	0.667	0.666	0.666
0.625	0.25	0.583	0.75	0.75
0.60	0.20	0.533	0.70	0.80
0.55	0.10	0.433	0.60	0.70
0.50	0.00	0.333	0.50	0.60
0.45		0.233	0.40	0.50
0.40		0.133	0.30	0.40
0.35		0.033	0.20	0.30
0.30		0.00	0.10	0.20
0.25			0.00	0.10
0.20				0.00

(二)不考虑猜测因素时易度与极大区分度的关系

从理论上讲,只要是固定选项个数选择型项目,无论是多选一,还是多选多,测量结果都不可能不存在猜测成分,项目的区分度也不可能不受到影响。换句话说,只有在项目是非

选择型项目或者不定项选择型项目的条件下，在对项目进行分析时才可以不考虑猜测因素。我们之所以计算出不考虑猜测因素时难度与其对应的极大区分度并列于表 2，主要是为了和考虑猜测因素时难度与其对应的极大区分度加以比较，从而揭示选择型项目选项数目与区分度的关系。不考虑猜测因素，并不意味着我们就不承认或者忽视猜测因素，而是我们假定猜测因素①为零。这是理论上的抽象和假定，在理论建设中是非常必要的，不可缺少的。

在不考虑猜测因素时，关于给定易度对应的极大区分度的计算方法与考虑猜测因素时的情况基本相同，惟一区别是，我们假定 $A=\infty$ 计算结果见表 2。

表 2 不考虑猜测因素时难度与极大区分度之间的关系

易度指数值 FI	极大区分度 $A=2DI_{MAX}$	极大区分度 $A=3DI_{MAX}$	极大区分度 $A=4DI_{MAX}$	极大区分度 $A=5DI_{MAX}$
1.00	0.00	0.00	0.00	0.00
0.95	1.00	1.00	1.00	1.00
0.90	0.20	0.20	0.20	0.20
0.85	0.30	0.30	0.30	0.30
0.80	0.40	0.40	0.40	0.40
0.75	0.50	0.50	0.50	0.50
0.70	0.60	0.60	0.60	0.60
0.667	0.666	0.666	0.6156	0.666
0.625	0.75	0.75	0.75	0.75
0.60	0.80	0.80	0.80	0.80
0.55	0.90	0.90	0.90	0.90
0.50	1.00	1.00	1.00	1.00
0.45	0.90	0.90	0.90	0.90
0.40	0.80	0.80	0.80	0.80
0.35	0.70	0.70	0.70	0.70
0.30	0.60	0.60	0.60	0.60
0.25	0.50	0.50	0.50	0.50
0.20	0.40	0.40	0.40	0.40
0.15	0.30	0.30	0.30	0.30
0.10	0.20	0.20	0.20	0.20
0.05	0.10	0.10	0.10	0.10
0.00	0.00	0.00	0.00	0.00

(三) 有无猜测因素存在时的对比分析

对比表 1 和表 2，我们不难看出，由于猜测因素的介入，即使在项目难度不变的情况下，项目的极大区分度在一定的范围内^⑧打了折扣，限制了项目区分度的挖掘潜力。而且我们还可以看出，猜测因素对于区分度的影响发生在极大区分度达到相应的最大极大区分度之后（沿易度指数值由大而小或自上而下），而在极大区分度还没有达到最大极大区分度之前，极大区分度并不受猜测因素的影响。对于以选拔或者筛选为目的的测试来说，区分度是试卷的

决定性质量指标,因此,构卷时(如果采用的是选择型题)就应该视具体情况适当降低项目的难度(即提高项目的易度)。我们通常说,当项目的易度指数值为 0.50 时项目的区分度极大,如果我们该文的研究结果可信,这一说法就不可一概而论。首先我们应该明确,心理测量学中的区分度有多种含义:有时相当于“区分力”(power of discrimination);有时相当于“灵敏度”(sensitivity);有时相当于别的,不好一概而论。

以上论断中的“区分度”是方差大小意义上的区分度,所谓的“当项目的易度指数值为 0.50 时项目的区分度极大”也是根据方差的定义确定的(参见谢小庆,1988:88-90;余嘉元,1987:167-169),这里的“区分度”就其本质而言应该是“区分力”。区分力大,仅仅只能保证测量的经济性,并不能保证测量结果区分的可靠性强、准确性高。而且,用区分力刻划的区分度,无论有无猜测成分,其值不变,即区分力与可靠度及准确度分家。不证自明,可靠性不强、准确性不高而仅仅是经济的测量是无用的测量,甚至是不公平或者有害的测量。这也是我们本文放弃采用“区分力”意义上的区分度,而采用“灵敏度”意义上的区分度的原因。灵敏度意义上的区分度和测量的可靠性以及准确性是一致的。也就是说,对于灵敏度意义上的区分度,只要区分度大,可靠性和准确性就不可能不高。

我们认为,就同样大小方差的测量结果(即同等区分力),由于猜测因素的介入,测量的误差必然会加大(有关测量误差方差的刻划和猜测误差的刻划问题,请参见席仲恩,2000a)。如表 1 所示,对于易度指数值为 0.60 的一个项目,其测量结果的区分力是不变的,但对于不同大小的猜测成分,实际可能引起的极大区分度是很不相同的。对于 $A=2$ 时,实际可能引起的极大区分度是 0.20;对于 $A=3$ 时,实际可能引起的极大区分度是 0.533;对于 $A=4$ 时,实际可能引起的极大区分度是 0.70;对于 $A=5$ 时,实际可能引起的极大区分度是 0.80。而如果不考虑猜测因素,则可能做出在以上四种不同情况下,区分度都为 0.80(见表 2)的决策,这显然是欠科学的。

四、模型的验证

为了验证我们的模型,我们统计分析了 1996 年~2000 年我国硕士研究生英语入学考试的试卷,发现我们模型的预测准确度相当高。五套考卷共 350 道四选一型客观考题,只有 60 道在预测范围之外。而且这 60 道题目大部分都是稍微超出范围,因为我们假定受考的水平是理想的正态分布,而且测量误差不存在,但实际上这种理想情况是没有的,所以这种小小的偏差是允许的。其他几道偏离较大的题目,我们怀疑是统计上的失误或者是原始数据印刷上的错误。例如,1997 年的第 25 道题,其易度指数值为 0.32,而区分度却为 0.86。

此外,我们还用周越美(Zhou Yuemei, 2003)的数据验证了模型,同样发现模型的与测性很好。在大学英语六级考试的 70 道题目中,仅有 10 道稍微超出范围。可见,这个模型的效度是很高的,其实用性也是显见的。

五、结束语

以上我们定量刻划了项目难度与极大区分度之间的函数关系,根据我们建立的数学模

型，列出了两个表格以供参考，我们还讨论了考虑猜测因素和不考虑猜测因素时同一难度项目的真正区分度在不同程度猜测因素介入时的差异。需要指出的是，我们提供的，仅仅是我们工作的方向和理想状况下所可能达到的极限。在实际工作中，我们认为应该最少注意两个方面的问题：一个是开发的效率问题，另一个是分清两类不同性质的区分度。

实际工作过程是很复杂的。由于各种原因，不可能每个项目的区分度都能开发到极限。我们应该认为，每个项目的区分度是不可能开发到极限的。一个项目的区分度潜力如果能被挖掘到 80%，就已经算很好了。因此，开发试卷时不可不考虑效率问题。假定我们认为我们的项目的区分度不可低于 0.40^①，那么选择项目时(假定 A=4)，就应该选择易度指数不低于 0.50、又不高于 0.75，极大区分度不低于 0.50 的项目，因为我们还要考虑区分度的挖掘效率。

关于两类不同性质的区分度，我们指的是建立题库时给每个项目标定的区分度和构卷时所选项目对于特定受考群的区分度。前者我们不妨叫它项目的固有区分度，后者叫它项目的特有区分度。我们认为：项目的固有区分度是合格项目的根本标志，是独立于项目相对难度(即相对于特定受考团体的项目难度)的；项目的特有区分度是合适项目的根本标志，是依赖于项目的相对难度的；合格是合适的前提。为了开发出合格的项目入库，建立题库时(主要指为选拔性考试而建立的题库)，用于实验项目的受考群的水平跨度应该尽量大一些，项目的区分度标准也要定得高一些，而且要基本一样。因为只有这样，用同一题库构出的不同试卷，其测量结果才便于等值^②(例如根据项目反应理论的单参数模型等值才有效)。只有在合格题库的前提下，根据特定受考构卷时只考虑相对于待考受考的项目难度就可以了，而且，只要难度合适，特定区分度也决不会低。此外，特定区分度和区分力、可靠性以及准确性也变得统一和同一。

最后我们要说明的是，这个模型刻划的是一种最大的可能性关系，并不是必然性关系。换句话说，一个项目的易度一定后，其对于特定团体的区分度最大可能是这么大，但是并不能保证就会这么大，而且很可能非常小。为了便于实践者参考，我们已经制作了项目易度与项目极大区分度之间关系的详细数表，需要者可向我们索取。

① 难度也可以直接用难度指数表示。常用的难度指数有 q 互补难度，ETS 难度，瑞查(Rasch)难度和席仲恩 H 难度。

② 由于猜测因素的存在，我们假定项目的易度指标不小于猜测概率。

③ 这个范围是选项个数的函数，选项个数越多(即猜测成分越小)，区分度的范围也越大。反之亦然。

④ 顺便举的例子。在实践中，最低区分度的确定取决于具体考试的重要性。

⑤ 根据朱正才(1997)和杨惠中、Weir(1998)，我国大学英语四、六级考试根据项目反应理论单参数模型等值，但在我们不能保证项目区分度基本一样的情况下是不能用单参数模型等值的，况且标准参照考试也没有必要在不同次考试之间等值。标准参照性考试的关键是每次考试的试卷是代表

标准(例如教学大纲)的样本。

参考文献:

- [1] Alderson, J. Charles, Clapham, Caroline, & Wall, Dianne. 1995. Language Test Construction and Evaluation. Cambridge: Cambridge University Press.
- [2] Anastasi, Anne. 1982. Psychological Testing(5th edition). New York: Macmillan Publishing Co. Inc.
- [3] Bachman, Lyle F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.
- [4] Bachman, Lyle F., & Palmer, Adrian S. 1996. Language Testing in Practice. Oxford: Oxford University Press.
- [5] Baker, David. 1989. Language Testing-A Critical Survey and Practical Guide. London: Edward Arnold.
- [6] Harris, David, P. 1969. Testing English as a Second Language. New York / St Louis / San Francisco: McGraw-Hill Book Company.
- [7] Heaton, J. B. 1988. Writing English Language Tests(2nd edition). Longman Group UK Limited.
- [8] Lado, Robert. 1961. Language Testing. London, Longmans. Green And Co Ltd.
- [9] Madsen, Harold S. 1983. Techniques in Testing. Oxford: Oxford University Press.
- [10] Salvia, John & Ysseldyke, James E. 1995. Assessment (6th edition). Boston / Illinois / Princeton: Houghton, Mifflin Company.
- [11] Spolsky, Bernard. 1995. Measured Words. Oxford: Oxford University Press.
- [12] Zhou, Yuemei(周越美). 2003. The Comparability Study of Two EFL Tests (CET-6 and TEM-4) in China. Unpublished Ph. D Dissertation, Shanghai International Studies University
- [13] 戴忠恒. 心理与教育测量[M]. 上海: 华东师范大学出版社, 1987。
- [14] 高兰生, 陈辉岳. 英语测试论[M]. 南宁: 广西教育出版社, 1996。
- [15] 桂诗春. 标准化测试一理论、原则与方法[M]. 广州: 广东高等教育出版社, 1986。
- [16] 教育部考试中心. 全国硕士研究生入学英语考试分析(非英语专业)[M]. 北京: 高等教育出版社, 2001。
- [17] 李筱菊. 语言测试科学与艺术[M]. 长沙: 湖南教育出版社, 1997。
- [18] 刘润清、韩宝成. 语言测试和它的方法[M]. 北京: 外语教学与研究出版社, 2000。
- [19] 舒运祥. 外语测试的理论与方法[M]. 上海 / 西安 / 北京 / 广州: 世界图书出版公

司, 1999。

- [20]王孝玲. 教育测量学[M]. 上海: 华东师范大学出版社, 1989。
- [21]席仲恩. 测试中的误差与测量结果的解释. 中国第八届当代语言研讨会, 广州, 2000a。
- [22]席仲恩. 扣分公式及其应用[A]. 见: 大连外国语学院学报编辑部(编). 外语研究与教学 (下册)[C]. 北京: 航空工业出版社, 2000b: P218~226。
- [23]席仲恩. 项目特征函数的导出及其特征研究[J]. 绍兴文理学院学报(自然科学版), 2001, (1)。
- [24]谢小庆. 心理测量学讲义[M]. 武汉: 华中师范大学出版社, 1988。
- [25]徐强. 英语测试的理论与命题实践[M]. 合肥: 安徽教育出版社, 1992。
- [26]杨惠中、Weir, C. 大学英语四、六级考试效度研究[M]. 上海: 上海外语教育出版社, 1998。
- [27]杨钟琳. 英语测试的原则与方法[M]. 杭州: 杭州大学出版社, 1992。
- [28]余嘉元. 教育和心理测量[M]. 南京: 江苏教育出版社, 1987。
- [29]于信风. 考试学引论[M]. 沈阳: 辽宁人民出版社, 1987。
- [30]朱正才. 大学英语四级考试(CET4)通过标准的制定[A]. 见: 张厚粲(主编). 心理与教育测量. 杭州: 浙江教育出版社, 1997: P278~284。