条目反应理论简述

刘仁刚
(深圳市康宁医院,广东 深圳 518020)

【摘要】 条目反应理论认为,每个测验条目有其自身固有的特征,比如难度、区分度等,这些特征与样本无关,也与其它条目无关。不同的潜在特质,适用于不同特性的条目来测量,这将提高测量精度和测验效率。条目的得分概率与潜在特质呈特定的回归关系,这些关系通常称为各种数学模型。本文还简要讨论了运用条目反应理论的注意事项。
【关键词】 条目反应理论; 心理测验; 综述

中图分类号: R395.1 文献标识码: A 文章编号: 1005-3611(2009)01-0037-05

A Brief Introduction to Item Response Theory

LIU Ren-gang
Shenzhen Kangning Hospital, Shenzhen 518020, China

【Abstract】 Item Response Theory takes the hypothesis that every test item has its own features such as difficulty, discrimination and others, which are not related to the sample and neither to other items. A given level of latent trait could be measured by specially featured items, so that the measurement accuracy and test efficacy will be increased. There is a certain regression between the score probability of an item and the latent trait, and the regressions are usually represented as mathematic models. This paper provided some brief discussions about the application of item response theory also.
【Key words】 Item response theory; Psychological test; Review

条目反应理论(Item Response Theory, IRT),也称条目特征曲线(Item Characteristic Curve, ICC)理论、潜在特质(Latent Trait)理论、条目反应模型(Item Response Model, IRM)、个人-条目反应理论(Person-Item Response Theory)。国内多数人将 Item Response Theory 翻译成"项目反应理论",作者参考了大量的国外文献,认为,Item Response Theory 中的 Item 一词的含义是明确的, 指的就是测验的单个条目或试卷的单个试题,在 IRT 这个短语以外的与心理测验或心理测量有关的应用中,item 的含义也是这样的,因此,作者使用"条目反应理论"这一中文名称。

条目反应理论的基本理念是每个条目有着自身固有的特性,每个条目都与被试的待测心理属性之间有着特定的函数关系,即,测验条目的得分概率(或通过概率)与待测属性之间存在确定的关系并可用数学模型来描述,只要知道了条目的特性,就可以估计被试的待测心理属性的值,而不必关心总体在该待测属性上的状态,例如总体的平均水平置信区间等等;而经典测验理论(Classical Test Theory, CTT)则认为待测心理属性须由一组条目来评估, 这一组条目称为行为样本,在对个体进行评估之前,必须用该行为样本测试相应总体的一个代表性样本,并获得常模数据,然后用这组行为样本测试被试,将被试的测验结果与常模进行比较来反映被试的待测属性的大小或强弱。在经典测验理论中,从总体抽取不同的样本,就会得到不同的常模数据,从而,单个被试的结果也会因常模数据的不同而不同。虽然经典测验理论也研究条目的特性,但并不特别看重单个条目与待测属性的函数关系。

我们可以按下述方式理解条目反应理论的上述基本理念。对于条目 i,某被试的待测属性越低,他在该条目 i 上得分的可能性就越低,反则反之。当待测属性的值处于某个区间时,条目 i 的得分概率将随着待测属性值的变化而快速变化,这说明条目 i 适合于测量待测属性处于这个区间的被试。而在另一些区间里,虽然待测属性有较大的变化,但条目 i 的得分率却保持不变或者变化不大。例如,对于某种待测能力 θ,假如将 θ 分为三个区域,第一个区域为负无穷大到 A,第二个区域为 A 到 B 且 B 大于 A,第三个区域为 B 到正无穷大,就所有 θ 低于 A 的被试来讲,条目 i 太难了,虽然这些被试在 θ 上的差异是很大的, 但他们在条目 i 上的得分机率却相似地低,条目 i 无法区分这些被试在该待测能力上的差异,这样,条目 i 就不适合于这些被试。换句话说,他们需要一些更容易的条目来区分他们在 θ 上的差异。相反,另一种情形是,对于所有待测能力 θ 高于 B 的被试来讲,条目 i 又太容易了,虽然这些被试在 θ 上的差异也是显著的, 但他们在条目 i 上的得分机率却相似地高, 条目 i 也无法区分这些被试在该待测能力 θ 上的差异。相反地,他们需要一些更难的条目来区分他们在 θ 上的差异。条目反应理论认为,人们可以找到足够多的难易不同的条目,根据被试在 θ 上的不同水平,选取最合适的条目实施测验,从而能够最有效地测量出被试的待测心理属性。我们可以用这样一个不太确切的比喻从某个角度来说明条目反应理论的基本理念:假如有了足够多的不同质量的砝码,我们就能够精确测量所有物体的质量,而不必事先了解所有物体的总体状况。

1 条目反应理论的内容

1.1 简要发展过程及相关概念

一般学者认为,条目反应理论的基本思想可以追溯到上世纪四十年代 Lawley 的工作[1],Lawley 认为,在一个给定的测验中,对条目的反应反映了某种潜在的特质(underlying trait)

或能力。1950 年,Lazarsfeld 第一次提出 “潜在特质”(latent trait)概念[2],这一概念一直沿用至今。因为迄今为止,所有心理属性都不能直接测量,所以,作者认为,潜在特质这一概念与条目反应理论没有必然联系, 使用待测心理属性亦无不可,本文在“潜在特质”和“待测属性”这两个名称的使用上不作区分。1952 年,Lord 在他的博士论文 “A theory of testing scores”中,系统地阐述了条目特征曲线理论[3],后来他将条目特征曲线理论改称为条目反应理论。这标志着条目反应理论的正式产生。1955 年,他发表文章对其理论做出进一步阐述[4]。在随后的 10 年中,包括 Lord 本人在内,美国心理测验界在这方面少有进展,主要原因可能是因为该理论的假设没有得到合理的验证。1965 年,Lord 通过一项对象达 10 万人的大规模研究才对其理论假设做出了初步的验证:该研究表明条目通过率(得分概率)与测验分数之间存在正态拱形(Ogive,亦译奥吉夫,或卵形)回归关系,而且条目的两个重要特征难度系数和区分度系数与样本无关(Sample Invariance),从而可以认为条目特征曲线理论的假设是成立的[5-7]。

1960 年,丹麦数学家 Rasch 独立提出了含有一个参数的“样本无关”的心理测验模型[8],这就是后来著名的 Rasch 模型,得到了较广泛的应用和较大的发展[9-12]。所谓样本无关是指条目的属性,如难度、区分度、猜测系数等与获得这些数据的样本无关,他们仅与条目本身有关。这些数据就是条目得分率与测验分数回归模型的参数或简称为条目的参数。Rasch 模型只含有一个参数,即条目的难度系数。可以这样理解样本无关的概念:研究者从总体中抽取样本来估计条目的参数,条目反应理论认为,该样本对总体的代表性如何关系不大,例如,样本的均数可能显著低于总体的均数等,只要在待测心理属性的每一水平上都包含了足够多的个体就可以了。

1968 年,在 Lord 和 Novick 合著的“Statistical theories of mental test scores”[13]一书中,著名的统计学家 Burnham 用四章的篇幅详细地阐述了条目反应理论中二参数、三参数的正态 Ogive 模型和 Logistic 模型的数学问题。至此,条目反应理论体系基本形成。

1969 年,Wright 和 Panchapakesan 提出了 Rasch 单参数模型的参数估计方法, 并编写了相应的计算机程序 BICAL。BICAL 的出现, 使条目反应理论的应用成为现实。此后,Wright 和其他许多统计学者或/和数学家不断努力,陆续提出了各种模型的参数估计方法[14]。在美国,参数估计的主流计算机程序是 BICAL、LOGIST、BILOG 和 FastTEST 等。

随着计算机运用的普及和计算能力的提高,条目反应理论的运用也越来越广泛,其数学模型也越来越复杂。有时,条目反应理论近乎成了数理统计学家或数学家的概率或数学游戏,有点脱离实际运用了。

1.2　基本内容

1.2.1　*条目反应理论的模型*　作者认为,到目前为止,条目反应理论的基本内容可以用等式(1)来概括:

$$P(_{i=j}|\theta) = f(a_i, b_i, ..., \theta) \quad (1)$$

在等式左边的表达式中,i 是指第 i 个条目,j 是指第 i 个条目的第 j 个答案,P 代表概率,指被试在第 i 个条目上回答 j 的概率,如果 j 是得分答案,那么 P 就是得分概率。在等式右边的表达式中,θ 指被试的潜在特质,亦即尚未明确的待测心理属性。在 IRT 的研究和运用中,多指潜在的能力或可能获得的成就,如果条目 i 与多个潜在特质有关,这里的 θ 就是多个,每增加一个 θ,$f(a_i, b_i, ..., \theta)$ 的复杂性就呈几何级数地增加许多倍,θ 的个数称为潜在特质的维度数, 在 IRT 发展之初,IRT 有一个假设, 即测验所测的被试的潜在特质是单维的,那么函数中就只有一个 θ,单维性假设虽然有利于理论阐述和数学计算,但却明显不符合实际,因为人们从测验实践中明确得知,所有待测心理属性均无例外地受各种因素的影响,被试对条目的反应不可能只受所假定的“单一潜在特质”的作用。

$f(a_i, b_i, ..., \theta)$ 是 θ 的函数(有时是积分函数),a_i 等是函数的未知的参数。例如,$f(a_i, b_i, ..., \theta)$ 可以是 $a_i\theta + b_i$,也可以是 $\frac{a_i e^{-\theta}}{1 + a_i e^{-\theta}}$,取什么样的函数依条目得分率(正确回答的概率)与 θ 的回归关系的类型而定,当然,所有这些函数都应是某种概率密度函数或者某种函数的积分函数, 这样,才能与等式左边的表达式相对应。常见的 $f(a_i, b_i, ..., \theta)$ 函数有正态分布的概率密度函数 $\frac{1}{\sqrt{2\pi}a_i} e^{-\frac{1}{2a_i^2}(\theta - b_i)^2}$、Logistic 分布的概率密度函数 $c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$ [15]等。

总结上面的内容,可以将等式(1)用文字表达为:潜在特质为 θ 的被试,在条目 i 上回答 j 的概率 P 是 θ 的函数,即条目 i 与潜在特质 θ 之间存在某种回归关系, 该函数受条目 i 的特性所制约。

不同的等式代表了不同的 IRT 模型。每一个等式或模型都可以描绘成一条曲线,这条曲线就称为条目特征曲线。常用的曲线有正态 Ogive 曲线和 Logistic 曲线,这两种曲线都与标准正态分布累积概率曲线非常相似,只是它们各有自己的参数,所以在形态和坐标系的位置上才与标准正态分布曲线有所不同。

理论上讲,IRT 模型是无数的。按照参数的多少,IRT 模型可以分为单参数模型、双参数模型直至四参数模型等;按照条目计分方式不同,可以分为两分法计分模型和多分法计分模型,而多分法计分模型又可以按条目答案有无相关性分为称名模型(条目的答案之间不相关)和等级计分模型;按照所测量的潜在特质的数目, 可以分为单维模型和多维模型,等等。顾海根等列举了基本上目前全部可用的具体的 IRT 模型[16],有兴趣的读者可以参考。

1.2.2　*信息函数*　信息函数给出了条目反应理论中测量的有效性度量,称为信息量。信息量等于标准误平方的倒数,即信息量越大,标准误越小,测量越精确。信息函数分条目信息函数和测验信息函数。

条目信息函数提供了单个条目的特征和潜在特质综合的信息量,在潜在特质的取值范围内,信息函数所计算的信息量是不均匀的。如果条目有难度参数,那么,当 θ 值在难度参数附近时,信息量最大,越往两端信息量越小。这是容易理解的,因为 θ 值越远离条目的难度,测验的误差就越大,条目

所能提供的准确信息就越少。条目的其它参数的情形与此类似。

测验信息函数估计了整个测验的精度，即测验信息量。测验信息函数等于全部条目的信息函数之和。测验的标准误等于测验信息量平方根的倒数。

信息函数是 IRT 的优点之一，它为不同特质水平的个体分别提供了测量精度。

1.3 理论假设

如果想要等式(1)成立，必须具备相当严格的条件，这些就是 IRT 理论的假设。在运用 IRT 理论编制测验条目时，不必过分理会这些假设，重要的是条目的得分率与测验分数之间存在可接受的回归关系，并且所选择的模型(即等式(1)的具体形式)能够在可接受的程度上反映这种回归关系。IRT 的各种理论假设相互联系，互为基础或因果，因此，应该综合考虑这些假设是否成立，如果不成立(严格意义上讲，不成立是肯定的)，就要考虑对假设的违反程度是否在可接受范围。

这些理论假设并不能通过有限的研究去证实它们普遍成立。原因之一，我们不能直接观测心理属性，所谓的潜在特质或待测心理属性都是对心理活动某一方面的概念化，并非心理活动的真实。例如，智力，它不过是对某种心理属性的抽象，不同的观察方法形成了不同的操作性定义，真实的智力，假如它的确存在的话，正好是我们所定义的吗？显然不是，我们只能从某个侧面近似地抽象它，因此，我们就有了多种关于智力的操作性定义。原因之二，我们有理由认为人的心理属性可能是无限多样的，每一种心理属性也可能有多种形态，这就使得我们很难用数学模型去精确描述它们。Lord 在那个有 10 万对象的研究中证实了部分假设的一般性成立，但是，他的研究结果并没有理由推广到所有心理属性。所以，我们只有在实际运用中，针对具体的待测心理属性，从直观经验和统计推断两方面检验所用的数学模型的合理性。

1.3.1 条目反应-潜在特质关联假设 IRT 假定，每个条目都与每个被试的所要测量的潜在特质密切关联，且符合相应的数学模型所表述的反应与潜在特质之间的关系，即，当被试回答正确时，的确表明了他是知道正确答案的，而不是因为其它原因使他回答正确；当被试回答错误时，的确表明了他是不知道正确答案的，而不是因为其它原因使他回答错误。其实，这是所有心理测验的基本假设。不过，经典测验理论在该假设上相对较弱，而条目反应理论在该假设上很强。CTT 从整个行为样本上假设一组条目与所要测量的特质相关联，在这一组条目中，对于某些被试来讲，可以存在个别或少数条目与待测心理属性不相关联的情形，允许有一定的误差存在，但是，IRT 则假定每个条目与每个被试的潜在特质相关联，且关联的方式和程度符合数学模型的描述。

目前，计算机化自适应测验(Computerized Adaptive Test, CAT)越来越普遍，在实施 CAT 时，计算机自动根据被试已经回答的非常有限的条目的答案来选择下一步要测验的条目，而且，整个测验的条目总数比较少，加之各个被试所接受的条目数不同，这样，满足该项假设就显得非常重要。如果被试的反应是因为疲劳、看错题、漏做题、随机猜测或胡乱作答、幸运地知道某个条目的正确答案等，那么，有限的一组条目所得的测验结果就反映不了潜在特质。

1.3.2 对象总体的潜在特质和条目的得分概率均为连续增分布 该假设的含义是：θ 以及函数 $f(a_i, b_i, \ldots, \theta)$ 的分布是连续且单调上升的；或者，如果 θ 为多个，则等式(1)所描述的曲面或超曲面在每个维度上都是连续且单调上升的。这一假设的意义是显而易见的。如果 $f(a_i, b_i, \ldots, \theta)$ 的值或者不是连续的或者不是单调上升的，那么，就会出现一个 θ 对应多个函数值(即多个得分概率)的情形，或者出现没有对应的函数值的情形，条目的得分概率与潜在特质就不存在一一对应关系，也就无法从得分率来估计潜在特质的大小。

1.3.3 潜在特质的大小在测验时刻是确定的 很明显，如果个体的潜在特质在测验时既可以是 A 也可以是 B，那么，等式(1)就根本不可能得到。物理学家已经发现，某些粒子在同一时刻可以出现在不同的地方，这些粒子在确定的时刻没有确定的位置，只能用概率来表示它们可能的位置。那么，心理属性是否也是这样？我们没有科学的答案。所以，如果要等式(1)成立，就只能假定心理属性在确定的时刻其大小是确定的。

1.3.4 局部独立性 局部独立性假设的内容是：被试对条目的反应除了与潜在特质有关外，不受其它条目的影响。如果条目 i 为条目 j 提供暗示或其它影响答案的因素，局部独立性假设就受到了破坏。其结果就是，条目 j 的难度系数等参数就不真实了。

1.3.5 维度空间 该假设的内容是：测验所测量的是模型所描述的一种或多种潜在特质的量，与其它心理特质无关，这些潜在特质构成测量的维度空间。如果存在模型以外的心理属性影响着模型中所描述的潜在特质，那么，这一模型就没有可靠性。这一假设对于模型的成立是基本的。最初的 IRT 理论为了表述和计算上的方便，假设测验所测的潜在特质是一维的，这自然引起强烈的反对意见。后来，统计和教育测量学家们将维度推广到二维甚至更多，但问题依然存在。很显然，将维度空间固定的假设是不成立的，且问题连环出现。如果维度数过少，很明显，有些潜在特质没有包括进来，且计算和结果解释也不明晰；如果维度数过多，那么，其中是否每个维度都是真实的呢？是否有些维度仅是一种假象甚至错误的呢？用什么工具，又需要多大的样本来检验这些维度的真实性和大小呢？作者认为，所有的心理测验都或多或少地违背了维度空间假设，只有脱离实际的理论家才会坚持认为他的模型符合他的维度空间假设。

从探索性主成分分析(或因素分析)到验证性主成分分析，从直线相关分析到曲线拟合优度的检验(这些方面的文献太多，为节省篇幅就不列举了)，研究者始终只能在某种程度上判断成分(这里就是维度)的多少及负荷的大小，总是有一定比例的测验结果不能用所选定的维度来分析。

那么，如何解决这一问题，并使 IRT 能真正应用于实际呢？作者认为，我们得回到传统测验理论的标准化问题上来。首先，不必过分依赖“潜在特质”这一概念。在选择条目时要遵循一定的原则，尽量寻找测量相同心理属性的条目。如果

发现一个测验测量了多个心理属性，我们可以将该测验分成几个分测验，或者将这组条目分到不同或新的条目组（库）里，尽量使同一组条目测量单一的特质。

1.3.6　条目特征曲线的性质　以θ为横坐标，以函数$f(a_i,b_i,...,\theta)$的值（即式(1)的左侧）为纵坐标构成坐标系，条目特征曲线就是式(1)在坐标图上的曲线形式。如果测验的是多个潜在特质，那么，曲线变成相应的曲面或超曲面。条目特征曲线的性质是一个综合性的假设，它包括了 1.3.2 和 1.3.3。

另外，它还包括下述三个方面的假设：

A、潜在特质的取值范围为$(-\infty, +\infty)$。这里，θ 的取值为标准分数，所以有负值。该假设认为潜在特质可以无限小，或者说没有确定的下限，也可以无限大，或者说没有确定的上限。

B、当条目i的概率函数$f(a_i,b_i,..., \theta)$小于某一数值时 c_i，曲线的左端以 c_i 为下渐近线，即得分概率无限接近于 c_i 但不等于 c_i，曲线也不会在某处翘起。这一假设的含义是，无论被试的潜在特质 θ 多么小，虽然被试的得分概率在条目 i 上无限接近于 c_i 但不会低于 c_i。无限接近于 c_i 就使曲线不会翘起。c_i 等于或大于 0。如果等于 0，则表明潜在特质的最小值接近 0；如果大于 0，表明潜在特质的最小值接近 c_i。有一些因素使潜在特质不接近 0 而是一个较小的正数，比如，猜测因素就有可能使条目的得分概率大于 0。必须注意，c_i 是曲线的下渐进线，等同于 θ 为负无穷大时函数$f(a_i,b_i,..., \theta)$的极限。许多作者将曲线在竖轴（即$f(a_i,b_i,..., \theta)$的值）上的截距当成 c_i，这是不正确的。出现这种情况的原因是：在实际运用中，研究者通常取 θ 的范围为(-3,3)作图，并将竖轴画在 θ=-3 的地方，这样，有些作者就容易将曲线在竖轴上的截距当成 c_i。

C、当 θ 大于某一数值时，曲线的右端以 r_i 为上渐进线，即得分率无限接近于 r_i，但不等于 r_i，曲线也不会在某处凹陷。r_i 等于或小于 1。r_i 等于 1 时表明，θ 充分大的被试在条目 i 上的得分概率接近 1。但是，有时，有某些因素使 θ 充分大的被试在条目 i 上的得分概率不接近于 1 而是小于 1 的一个正数 r_i。比如这些被试可能考虑过多而在条目 i 上回答错误，此即 Hoffman 效应。Hoffman 效应由 Hoffman 于 1962 年提出。其内容大致是：能力高的人知道得太多，他们在对条目做出反应时考虑到的因素超出了条目本身所涉及的范围，这反而可能导致他们做出错误的反应。

能够基本满足上述六个假设的函数通常是正态累积概率函数和 Logistic 函数。因此，目前 IRT 数学模型也主要是这两种函数或它们的变型。

2　条目反应理论的运用

条目反应理论的运用分两个主要方面，第一，建立条目库，实施组卷测验或计算机化自适应测验；第二，运用条目反应理论对已有测验进行条目分析，并且也有可能对这些测验实施计算机化自适应测验。

2.1　建立条目库的一般过程

一般运用过程如下：①确定被试总体→②确定测验目标或内容→③根据测验目标初步筛选条目→④抽取足够大的样本，样本应涵盖 θ 的全部取值范围→⑤在样本中实施测验→⑥确定数学模型，使用计算机程序估计条目参数(具体方法参照程序的使用手册)→⑦根据条目参数分析条目的性质→⑧修改、删除质量差的条目→⑨重复⑤到⑧直到条目质量达到要求→⑩将条目加入到条目库。

2.2　条目分析

不管基于什么理论编制测验或条目，都可以运用 IRT 来分析条目的质量。具体的做法是：观察条目得分与总分的点阵图，试探性地确定数学模型，然后用相应的计算机程序分析原始答案矩阵，得出各条目的参数，这些参数就表明了条目质量的好坏。如果对所取的数学模型不满意，可以再试用其它数学模型。运用 IRT 分析条目质量只是条目分析的手段之一，并非所有测验的条目都适合于 IRT 分析。在选择数学模型时，参数的取舍也很重要，例如，对于不存在难度问题的测验，就不必在模型中加入难度系数。

2.3　计算机化自适应测验

计算机化自适应测验是 IRT 的典型应用。测验过程如下：计算机按某种规则 A 取出一个条目，根据被试的答案以及答案的信息量，选择下一个条目，如此循环进行，直到测验结果满足某个预先设置的规则 B。规则 A 和规则 B 都可以是多个。例如，规则 A 可以是难度系数居中或区分度系数最大；规则 B 可以是信息量达到某一个值或条目达到一定的数量。

2.4　组卷测验

在条目库中按一定的规则抽取固定数量的条目组成试卷实施测验。按测验的需要确定选题的规则。

3　运用条目反应理论的注意事项

3.1　心理属性的界定

如果给了你足够多的不同质量的砝码，你能测量出所有物体的质量吗？答案是否定的。当物体的质量大或者小到一定程度，利用杠杆原理测量物体的质量就不合适了。物理测量尚且如此，心理测量，它具有更大的不确定性，就更应该明确地界定测量的内容及范围。

心理属性和物理量没有可比性，例如，我们还无法得知，儿童的智力与成人的智力究竟是同一个心理属性在量上的不同，还是不同的心理属性，也就是说，儿童的智力和成人的智力可能根本就是不同的两种心理特质。因此，心理测验仍然要对被测属性作出较为可靠的实际可用的操作性定义或界定。

在什么时候，我们需要“用科学的方法”来了解一个小学三年级学生的数学水平和一个高中三年级学生的数学水平的差别呢？实际地讲，心理测验还是要针对特定的对象，还是要考虑到不同对象之间同一人为名称的属性（推测中的潜在特质，或人为的概念化的心理属性）可能代表着本质上十分不同的心理特征。

我们所概念化的潜在特质或待测心理属性仍然处于黑箱之中。首先它们是否就是我们所概念化的那样就值得怀疑；其次，这些特质或属性对刺激的反应远非像物体在天平上对砝码的作用那样直观、单纯和易于测量，所以，我们大可不必对我们提出的数学模型过于乐观。

3.2 选择合适的心理属性作为测量的内容

在 IRT 理论中,难度系数是条目的基本属性之一,目前较成熟的 IRT 模型都包括了难度系数。对于不考虑条目难度的测验,例如,在多数人格测验中,利用 IRT 理论来分析条目或编制和实施测验时恐怕主要应考虑区分度系数。

另外,至少到目前为止,条目本身的许多属性与待测心理属性之间的关系还不清楚。例如,在最高成就测验中,条目得分反应的概率与潜在特质相关,但是,在典型行为测验中,条目得分反应的概率与潜在特质的关系并不确定,只有一组条目的得分之和才有意义。再比如,许多异常心理的判断只需要有限个数的核心反应,可以引出核心反应的条目和只能引出非核心反应的条目与待测的异常心理的关系就大不相同。如何描述条目的这些性质仍有待于进一步研究。

3.3 全错与全对问题

IRT 的条目特征曲线假定,函数 $f(a_i,b_i,...,\theta)$ 最低以 0 为下渐近线,最高以 1 为上渐进线,因此,在估计条目特征时,如果有全错或全对的条目,那这些条目的特征参数是无法估计的。出现这种情况时要考虑下述可能:条目本身不合适,应当剔除;条目有漏洞,需要重新研究并加以修改;样本过少,样本没有包括各种水平的潜在特质的被试,这时就要加大样本,增加具有高特质水平或低特质水平的被试。

3.4 个体的测验分数是否需要与总体比较

如果个体的能力估计值只是与组内的其他个体相比较,可以首先考虑 IRT;如果个体的测验分数需要与总体比较,我们认为还是选用 CTT 为好。只是应该注意,在选用 CTT 时,当个体的分数与常模相差太远,比如超过 3 个标准差,我们就要了解到这时的实际标准误是比较大的,个体分数的可靠性降低,如果条件许可,应当选用更合适的测验。

3.5 条目间的顺序关系

IRT 使用者应该考虑条目在什么特性上有顺序关系以及这种顺序关系的意义是什么。

许多作者忽略了这一点,而这一点又是非常重要的,IRT 的多个假设都与此有关。如果假定了条目的得分概率与潜在特质之间存在确定的函数关系,而该函数又属于单调增函数,那么,条目之间就必定存在顺序关系。当然,从前面的讨论可以看出,这种顺序关系通常是难度方面的顺序关系。只有存在这种顺序关系,计算机化自适应测验才可能实现,IRT 的优点才可能体现出来。但是,许多心理测验所用的行为样本(即条目组)不存在或者根本不需要这种顺序关系。如果不考虑这一点,对那些不需要顺序关系的条目盲目计算出来的条目参数没有任何实际意义。

如果测验对条目没有难度要求,可以设定所有条目的难度系数都为零(即标准分数的均数),这时,条目可能在区分度参数方面存在顺序关系。这种情况下,IRT 可以利用这一顺序关系做三件事:一是修改或剔除区分度过低的条目,组成更有效率和更精简的测验;二是在分析测验结果时,对于得分相似的被试,那些在区分度高的条目上得分的被试的误差比在区分度低的条目上得分的被试的误差要小,通过计算信息量重新估计误差并调整置信区间;三是在选拔(或淘汰)测验中,如果只需要划界分,可以试行计算机自适应测验,这有可能提高测验精度或节约成本。

3.6 关于样本无关与参数不变

许多因素都能影响到样本无关(或参数不变)这个 IRT 的理想,譬如,样本没有包括所有 θ 水平的被试,潜在特质空间以外的特质对条目反应的影响,其它各种随机误差,等等。所以,实际上,任意两个样本所估计的条目参数都是不同的。如果所抽取的样本不能较好地代表总体,那么估计的条目参数就会出现线性偏移,不同的样本,偏移的方向和程度均不相同,这时需要对这些参数进行线性转换[17],才可以对两样本或样本之间的个体进行比较。

3.7 谨记条目反应理论的假设

这是一条不言自明的注意事项。如果观测数据不可接受地违反了 IRT 假设,那就不应采用 IRT 来设计测验或分析条目了。

4 小 结

IRT 萌芽于 40 年代,理论形成于 60 年代,70 年代后在教育心理测量方面得到广泛应用,近来在临床心理方面的应用也逐渐增加。IRT 的基本思想是每个测验条目有其自身固有的特性,例如难度、区分度、猜测度等,这些特性独立于样本,也与其它条目无关。条目的得分概率是被试潜在特质的函数,即条目的得分概率与被试的潜在特质有确定的回归关系,这些回归关系可以表达为各种各样的数学模型。条目反应理论可以运用于条目分析、计算机化自适应测验等。在运用 IRT 时,要注意它的理论假设和其它注意事项,避免运用不当。(致谢:本文得到了戴晓阳教授的有益指导,特表谢忱。本文的所有错误及遗漏之处,皆由作者自己负责。)

参 考 文 献

1 Lawley DN. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61:273-287
2 Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. In Stouffer L, Guttman EA, Suchman PF, et al. Studies in social psychology in World War II, Vol. IV: Measurement and prediction. Princeton, NJ: Princeton University Press, 1950.362-412
3 Lord FM. A theory of testing scores. Psychometric Monograph, 1952
4 Lord FM. Some perspectives on the attenuation paradox in test theory. Psychological Bulletin, 1955, 52(6):505-510
5 Lord FM. A strong true-score theory, with applications. Psychometrika, 1965, 30(3):239-270
6 Lord FM. A note on the normal ogive or logistic curve in item analysis. Psychometrika, 1965, 30(3):371-270
7 Lord FM. An empirical study of item-test regression. Psychometrika, 1965, 30(3):373-376
8 Rasch G. Probabilistic models for some intelligence and attainment tests.Copenhagen: Danish Institute for Educational Research, 1960

(下转第 50 页)

tional expression and localization. FEBS Lett, 1994, 341(1): 33–38
4 Witta J, Palkovits M, Rosenberger J, et al. Distribution of nociceptin/orphanin FQ in adult human brain. Brain Res, 2004, 997(1):24–29
5 Sumei Liu, Hong–Zhen Hu, Jun Rren, et al. Pre– and post–synaptic inhibition by nociceptin in guinea pig small intestinal myenteric plexus in vitro. Am J Physiol Gastrointest Liver Physiol, 2001, 281(1):237–246
6 Nothacker HP, Reinscheid RK, Mansour A, et al. Primary structure and tissue distribution of the orphanin FQ precursor. Proc Natl Acad Sci USA, 1996, 93(16):8677–8682
7 Berger B, Rothmaier AK, Franziska W, et al. Presynaptic opioid receptors on noradrenergic and serotonergic neurons in the human as compared to the rat neocortex. Br J Pharmacol, 2006, 148(6):795–806
8 Gyenge M, Hantos M, Laufer R, et al. Effect of nociceptin on histamine and serotonin release in the central nervous system. Acta Pharm Hung, 2006, 76(3):127–132
9 Tao R, Ma Z, Thakkar MM, et al. Nociceptin/orphanin FQ decreases serotonin efflux in the rat brain but in contrast to a kappa–opioid has no antagonistic effect on mu– opioid–induced increases in serotonin efflux. Neuroscience, 2007, 147(1):106–116
10 Connor M, Vaughan CW, Chieng B, et al. Nociceptin receptor coupling to a potassium conductance in rat locus coeruleus neurones in vitro. Br J Pharmacol, 1996, 119: 1614–1618
11 Marti M, Stocchi S, Paganini F, et al. Pharmacological profiles of presynaptic nociceptin/orphanin FQ receptors modulating 5–hydroxytryptamine and noradrenaline release in the rat neocortex. Br J Phamracol, 2003, 138(1):91–98
12 Liu Z, Wang Y, Zhang J et al. Orphanin FQ: An endogenous antagonist of rat brain dopamine transporter. Neuroreport, 2001, 12(4):699–702
13 Judd AK, Kaushanskaya A, Tuttle DJ, et a1. N–terminal modifications leading to peptide ORLI partial agonists and antagonists. J Pept Res, 2003, 62(5):19l–198
14 Kotlinska J, Rafalski P, Biala G,et al. Nociceptin inhibits acquisition of amphetamine–induced place preference and sensitization to stereotypy in rats. Eur J Pharmacol, 2003, 474(2–3):233–239
15 Zaveri N. Peptide and nonpeptide ligands for the nociceptin/orphanin FQ receptor ORL1: Research tools and potential therapeutic agents. Life Sci, 2003, 73(6):663–678
16 **沈渔邨,主编. 精神病学. 第四版. 北京: 人民卫生出版社**,2001. 426–454
17 Calo' G, Rizzi A, Rizzi D, et al. [Nphe^{1},Arg^{14},Lys^{15}]nociceptin–NH_2, a novel potent and selective antagonist of the nociceptin/orphanin FQ receptor. Br J Pharmacol, 2002, 136:303–311
18 Gavioli EC, Marzola G, Guerrini R, et al. Blockade of nociceptin/orphanin FQ–NOP receptor signaling produces antidepressant–like effects: pharmacological and genetic evidences from the mouse forced swimming test. Eur J Neurosci, 2003, 17:1987–1990
19 Calo' G, Guerrini R, Bigoni R, et al. Characterization of [Nphe^{1}]nociceptin(1 – 13)NH_2, a new selective nociceptin receptor antagonist. Br J Pharmacol, 2000, 129:1183–1193
20 Ozaki S, Kawamoto H, Itoh Y, et al. In vitro and in vivo pharmacological characterization of J–113397, a potent and selective non–peptidyl ORL1 receptor antagonist. Eur J Pharmacol, 2000, 402:45–53
21 Redrobe JP, Calo' G, Regoli D, et al. Nociceptin receptor antagonists display antidepressant–like properties in the mouse forced swimming test. Naunyn–Schmiedeberg's Arch Pharmacol, 2002, 365:164–167
22 **胡电,古航,熊英,等. 产后抑郁症与孤啡肽及单胺类递质的相关性研究. 中国神经精神疾病杂志**,2003,29(5): 321–322
23 **郑洪波,王斌,张璐璐,等. 抑郁症患者与健康人血浆孤啡肽含量的初步研究. 国际医药卫生导报**,2007,13(1):4–6
24 Stahl SM. Mixed depression and anxiety: Serotonin1A receptors as a common pharmacologic link. J Clin Psychiatry, 1997, 58:20–26

(**收稿日期**:2008–06–28)

(**上接第 41 页**)

9 Davision ML, Chen TH. Parameter invariance in the rasch model. The Annual Meeting of the American Educational Research Association, Chicago, 1991
10 Schumacker RE, Randall E. Rasch–based factor analysis of dichotomously scored item response data. The Annual Meeting of the American Educational Research Association, Chicago, 1991
11 Kyngdon A. The rasch model from the perspective of the representational theory of measurement. Theory and Psychology, 2008, 18(1):89–109
12 Borsboom D, Scholten AZ. The rasch model and conjoint measurement theory from the perspective of psychometrics. Theory and Psychology, 2008, 18(1):111–117
13 Lord FM, Novick MR. Statistical theories of mental test scores. Reading MA: Addison–Wesley, 1968
14 Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press, 1982
15 **郭庆科. 心理测验的原理与应用. 北京:人民军医出版社**, 2002. 98
16 **顾海根. 心理与教育测量. 北京:北京大学出版社**,2008. 122–128
17 Johnson W, Spinath F, Krueger RF, et al. Personality in Germany and Minnesota: An IRT–Based Comparison of MPQ Self–Reports. Journal of Personality, 2008, 76(3): 665–698

(**收稿日期**:2008–10–24)