简论测验等距假设

刘仁刚
（深圳市精神卫生研究所，深圳市康宁医院，深圳 518020）

【摘要】 多数心理测验只能达到顺序量表水平，不适合算术运算和参数统计检验，测验分数是心理属性的一个区间的代表，而不是一个点的精确表示。但是，在实际运用中，我们假定心理测验是等距量表，即，条目内、条目之间、被试之间的分数具有相同的单位。我们认为条目和对条目的反应具有质性和量性的区别，因此，提出了条目质性的概念；我们还认为，有的心理属性的行为样本不具备难度性质，或者条目的难度没有实际意义，或者有理由假定条目难度是相等的，我们提出条目集群的概念来处理这类条目的分数分配方法。经典测验理论的标准分和条目反应理论的logit是在等距假设前提下对测验分数的数学表示，虽然它们本身都是等距量表，但并不代表测量内容也是等距的。明确等距假设的风险有利于更好地运用测验。有多种途径可以降低等距假设的风险。

【关键词】 心理测验；等距假设；条目质性；条目集群

中图分类号：R395.1 **文献标识码**：A **文章编号**：1005-3611(2014)05-0845-04

A Brief Discussion about Test Equidistant Hypothesis

LIU Ren-gang
Mental Health Institute of Shenzhen, Corning Hospital, Shenzhen 518020, China

【Abstract】 Instead of inspection of arithmetic operation and parametric statistics, most psychological tests can only achieve the level of ordinal scale. The test score stands for one interval of a psychological attribute rather than precise expression. However, it is assumed that the scale of psychological tests is equidistant in practices, meaning the grades between items, between testees and within an item have the same unit. We suggest that there are qualitative and quantitative differences between items and the reactions to items. Hence the concept of item qualitativeness is raised. Besides, we think that some behavior samples of a psychological attribute have no difficulty, or the difficulty of items have no practical significance, or it is reasonable to assume that the difficulties of items are the same. The idea of item colony is proposed to handle the grading method of this kind of items. The standard score from classical test theory and the logit from item response theory are mathematical expressions of test scores under the assumption of equidistance. Although both of them are equidistant scales, it doesn't mean that the contents of the measurement are equidistant. Understanding the risks of equidistant hypothesis can better help test utilization. There are multiple paths to reduce the risks of equidistant hypothesis.

【Key words】 Psychological test; Equidistant hypothesis; Item qualitiveness; Item colony

1 相关概念与问题的提出

1.1 测验分数

统计学者认为，从测量学的角度来看，行为科学的测量大多数情况下充其量只能达到顺序量表水平[1]。临床心理学家也很清楚心理测验除了测量水平最高只能达到顺序量表水平外，它还具有另一个特殊性：自变量不是一个或几个处理因素，而是一组刺激，因变量不是一个或几个观测值，而是与这组刺激有关的一组反应，这组反应就是与待测验心理属性有关的全部行为的一个样本，被称为行为样本[2]。通常，将这组刺激称为一组条目(Item)，并将这组反应用分数表达出来。为了叙述方便，下面用表达式说明前述内容。设有一个测验T，共有I个条目(ITEM)，N个被试实施了测验T，得到N份答卷，与I个条目对应，每份答卷有就有I个分数，第j个被试的第k个条目的分数用$ITEM_{jk}$表示，j=1,..,N,k=1,..,I。用S表示这些答卷，我们就有如下分数距阵：

$$S=\begin{bmatrix} ITEM_{11} & \cdots & ITEM_{1I} \\ \cdots & \cdots & \cdots \\ ITEM_{N1} & \cdots & ITEM_{NI} \end{bmatrix}$$

S中的元素代表每个被试在每个条目上的分数。这些分数的意义不可能被赋与精确的数量含义，它们只是表示了分数之间的相对大小。我们并不知道这些数字之间的准确距离，或者说，我们不知道这些数字的单位，也就无法知道任意两个数字之间的真实距离，所以，这些数字最多只能是表示某种顺序。

下面以一个例子加以说明。龚耀先修订的艾森克个性测验(成人)的一个关于外向性的条目k：

你是否有许多不同的业余爱好？.....1是 2否

测验时，在外向性分量表上，该条目答1记1分，答2记0分。一组被试的答卷S就由0和1两个元素构成。在S中，我们不清楚任意两个数字(即使这两个数字相同)之间的距离，比如，设$ITEM_{jk}$和$ITEM_{(j+1)k}$都是1，表示被试j和被试j+1在上述条目上都答“是”，很清楚，虽然这两个被试都有许多不同的业余爱好，因此，都被赋与了相同的分数1，但是，他们在业余爱好的多少、兴趣程度、爱好的性质(如，是与生命体有关，还是与无生命体有关?)等等方面是不相同的，对于外向性这个心理属性的意义也就不同，这两个1也就不相等。

有些测验的条目有不同的等级或数字个数，例如，第 k 个条目的分数可能是0、1，而第 k+1 个条目的分数可能是1、2、3，如此等等。

1.2　心理属性的分布假设与分数的意义

那么，S中的数字的真实意义究竟是什么呢？在多数心理测验中，这些数字代表一个基础的连续增分布的一段区间。一般情况下，我们可以假定心理属性，包括个别行为，具备基础的连续增分布。以示例来说明就是，“有许多不同的业余爱好”具有一个基础的连续增分布，即，条目 k 从“极端否”到“极端是”对于外向性这个属性来讲是连续增加的。设计测验时，将这个连续增分布分为两段，靠近“极端否”记为0，靠近“极端是”记为1。所以，在记分为0的被试中，有较接近1的，在记分为1的被试中，有较接近0的。有的被试无法判断他自己是否有许多不同的业余爱好，可能随便回答，这时，他的分数的意义就更不能确定了。

综上所述，如果我们接受基础的连续增分布的假设，我们就可以确定S中的分数代表着一个区间，而不是一个点，每个数字在区间上的位置因不同的被试和/或不同的条目而不同，即使是相同的被试或/和条目，也因不同的时间和空间而不同。

1.3　等距假设的内容

事实很清楚，在S中，元素之间只有顺序关系。数据之间没有确定的单位，我们无法知道任意两数据之间差的精确值，因此，它们与算术不同构，不适合进行算术运算，也不适合运用参数统计。

Guilford和Fruchter明确知道测验只有顺序水平，但他们武断地认为可以假定达到了等距水平[3]，即S中的数字之间有相等的单位。这就是心理测验的等距假设(equal interval hypothese)。只有在这个假设基础上，我们才可以对分数进行数学运算。

在此基础上，经典测验理论(Classical Test Theory，CTT)将每个被试在各条目上的得分相加，得到每个被试的总分或因子分SCORE，此即测验的结果，设为R，N个被试的结果表示如下：

$$R = \begin{bmatrix} SCORE_1 \\ \cdots \\ SCORE_N \end{bmatrix}$$

相应地，测验等距假设也指R内的各分数具有相同的单位。

1.4　条目质性

这是一个与等距假设密切相关的问题，至今还没有看到有研究者认真讨论它。我们对条目的质性和量性的定义是：条目的质性是指条目分数反映了心理属性的质的不同；条目的量性是指条目的分数反映了心理属性的量的不同。我们希望质性和量性概念有助于处理测验等距假设问题，也希望为心理测验研究提供一个新的方向。和条目类似，有时一组条目也有质性和量性的区别。

我们有信心假设，多数条目同时具备质性和量性，但有明显的倾向性。少数条目基本上是质性的，另有少数条目基本上是量性的。

因此，等距假设的另一个内容是忽略了条目质性。

PHQ-9[4]是DSM-5推荐的抑郁严重程度的自评量表，它的9个条目完全符合DSM-5关于抑郁发作的9个症状。下面是两个示例条目：

1：做事时提不起劲或没有兴趣 0完全不会 1好几天 2超过一周 3几乎每天

3：入睡困难、睡不安稳或睡眠过多.. 0完全不会 1好几天 2超过一周 3几乎每天

评定时间是在过去的两周内，一天之中，症状持续存在，即使有波动，也不会根本消除。

第一条是抑郁发作的必备条件之一，如果回答3，则符合，否则不符合，而第三条则不是必备条件。因此，可以认为第1条为质性条目，“几乎每天”为质性反应，而第三条到第九条为量性条目。但是，作为一组相关的条目，第三到第九条有4条回答3，也具备诊断意义，因此，第三至第九条综合起来也是质性的。

上面的分析清楚表明，PHQ-9的分数分配不符合等距假设。

PHQ-9的作者将该量表用于抑郁严重程度的评估和抑郁发作的筛查。因为“抑郁严重程度”已经由量表来定义了，所以，信效度均高。筛查效果灵敏度和特异度均为88%(样本量580)。我们认为，如果分数分配更恰当些，筛查效果将更好，甚至可以用于初步诊断。

1.5　条目集群

从条目难度(或严重程度、频度等)这个意义上讲，等距假设的内容也应该包括对条目难度差异的忽略：事实上，部分心理属性的行为样本不具备有意义的难度，或者有理由假定条目难度是相等的。我们认为，心理障碍的非核心症状、智力中的常识等均不具备有意义的难度。对于这样的心理属性，应该以条目组为反应单位，我们将这组条目称为条目集群。我们以智力中的知识和精神病学上的症状条目为例，对条目集群的不同处理方法：对于等难度的知识条目，条目集群中一定数量的条目通过即为该条目集群在整体上通过，而不需要累加所有条目的通过数来比较智力的高低；而对于等难度的精神症状条目集群，则应该累计条目得分总数，以此反映症状的广泛性和严重性，以及，如果可能的话，反映“疑病”属性的大小。

2　等距假设的风险

虽然CTT实践已经证明，等距假设并没有阻止我们运用心理测验，但是，明确等距假设的风险有利于更加谨慎地更好地编制和使用测验，避免因使用不当带来的不良后果。

2.1　测量学风险

将顺序量表提高到等距量表并采用参数统计推断时，增加了犯第一类错误的风险，即，拒绝了本不该拒绝的零假设。一种典型的情形是有关药物疗效的研究。对治疗效果的评估所得到的分数不是等距量表，但是，多数研究者不加

考虑地按等距量表来使用参数检验,因为参数检验拒绝零假设的效率比较高,所以,这些研究所得出的拒绝零假设的结论是值得怀疑的。

2.2 错误分析测验内容

许多在同一名称下的心理属性,在不同量的情况下可能代表着不同质的内容,例如,艾森克个性测验里的外向性,等距假设让我们自然地以为20分的外向性和50分的外向性只是量上的区别,而忽略了20分的外向性分数可能已经不是正常的内向了,很有可能是另一种心理属性的表现,但被我们忽略了。

2.3 忽视了条目权重

等距假设容易让测验编制者忽略条目对分量表的贡献率和分量表(或条目组)对待测属性的贡献率,一个典型的例子是韦氏智力测验,各分量表以等量加权的方式加和成全量表粗分[5],忽略了每个分量表对智力的贡献率是不同的这一事实。

2.4 忽视群体的异质性

错误地将适用于某个群体的测验用于另一个完全异质的群体。我们知道,SCL-90适用于有心理精神症状的群体,并不适用于正常人群。由于等距假设,许多研究者便产生一个错觉,将SCL-90用于正常群体,只是分数低些而已。殊不知,这种低分说明了群体间质的不同。这是许多研究者不自觉犯的错误。王文中和吴齊殷[6]将SCL-90用于学生,发现绝大多数条目的记分都可以改为0、1。我们认为,如果他的研究再深入一些,会发现大多数条目都应该撤换。如果我们谨记心理测验的不等距性,就会少犯这类错误。

3 降低等距假设的风险

3.1 测验内容的定义

既然我们不能直接观测测验内容或潜在特质,那么,对测验内容的定义(概念化)就尤其重要。对测验内容的良好定义至少应该包括下述几个方面:①尽量完整地收集欲定义的测验内容应该有的和不应该有的行为表现,通过预试验来分析定义的质量并完善定义,这一过程可能需要重复多次;②具有良好的可操作性;③可验证性或可重复性,不仅指时间上的,有时也包括跨群体一致性;④尽可能明确测验内容在量上的不同是否反应了质的不同,例如,艾森克个性的P量表,作者在实践中发现,很多极端P分,并非手册所述的意义,相反,有的被试表现出高度成熟和智慧并有深刻的情感活动,有的则如精神分裂症患者一样情感淡漠。

心理测验是为了尽量客观地了解人的心理活动,因此,纯理论的、不可实证的、基于个案的概念是不适合测验的。

有把握认为,在今后相当长的时间内,我们还不能像了解物质的质量那样来了解心理活动的特征,所以,绝大多数概念化的心理活动都只有有限的时空有效性。

3.2 测验的目的或用途

不同目的的测验,对等距假设的处理有所不同。目前多用的有如下几类测验:①鉴别测验或诊断测验;②选拔测验;③水平测验;④效标测验;⑤预测测验。明确测验目的对编制高质量的测验也有重要意义。例如,通常,高考时,少数极难的题目反而分配较少的分数,原因在于,高考不完全是水平测验,更重要的目的是选拔适合上大学的考生,了解考生的学业能力或学业水平的整体分布只是次要目的。

3.3 合理地确定条目分数

确定条目分数包括一个条目应该分为多少等级及两个相邻的等级间的间隔大小。除了明确遵守前面几节的原则外,还应该重视下面几条。

3.3.1 *专家经验* 在我们还不能客观地测量心理属性之前,专家经验至关重要。例如,下面是SCL-90的抑郁量表的两个条目:14.感到自己的精力下降,活动减慢 1无 2轻度 3中度 4偏重 5严重;15.想结束自己的生命 1无 2轻度 3中度 4偏重 5严重。测验编制者应该认真分析这两个条目是否应该分配同样的分数,同一个条目内,不同等级间的差应该多大才合理,等等。

3.3.2 *检验量表分的分布* 对于心理属性,我们可以接受的一个假定是,在一个均质群体里,心理属性为正态分布或接近正态分布的其它分布。初步给定条目的分数后,再检验量表分是否符合正态分布,如果符合,进行条目分析,如果不符合,说明分数不合理或者不能假定待测心理属性符合正态分布。此外,还应该结合条目分析的结果来协助修正条目的分数。

3.4 条目分析

3.4.1 *条目数量* 每个测验内容都要有足够多的条目数量。S中的数字代表一个区间,我们不知道区间的单位,也就不知道区间的长度和数字之间的距离,更不知道条目得分分布中该区间下的面积。如果数字分配是合理的,那么,我们就有理由假定,该数字在区间中的位置是随机的,其误差符合随机误差的假定,按中心极限定理,足够多的条目的随机误差之和趋于0,即,条目分数的不等距不影响量表分数的等距假设,或者说,量表分数的等距性假设是可以接受的。如果因实际情况,量表的条目数达不到应有的数目,导致量表分数的分布不是基础的增分布,那么,我们就不应该按等距量表来处理量表分。

3.4.2 *条目的含义* 例如,在SCL-90的抑郁量表中有这样一个条目:14. 感到自己的精力下降,活动减慢 1无 2轻度 3中度 4偏重 5严重。我们应该怀疑,“精力下降”和“活动减慢”放在一起是否合适。

3.4.3 *条目难度的特殊问题* 我们在条目集群节讨论了等难度问题。目前最常用一个标准的等难度测验是二项必选数字记忆测验[7]。由于二项必选数字记忆测验的特殊性,测验计算了所有条目的部分。这个测验的条目适合用0和1分配分数,量表分为二项分布,按二项分布分析量表分就相当容易了。对于有难度的条目,CTT已经有了相当多的分析方法,此处不重复。近几十年来,条目反应理论(Item Response Theory,IRT或译项目反应理论,或译試題反應理論)对于条目分析也有了相当成熟的方法[6,8],此处亦不重复。

3.4.4 *条目质性* 条目质性的分析参见本文“条目质性”节。

4　关于等距的两个错误观念

S内各元素至多处于顺序量表水平，如果不能合理地假设它处于等距量表水平，或者不能合理地假设非等距属于随机误差，那么，基于S的任何算术运算都是不可靠的。即使满足前述条件能够将S施以算术运算，计算结果和心理属性仍然属于两个不同的范畴，不能把计算结果的算术特征反过来加在心理属性上。下面是两个常见的错误概念。

4.1　CTT的标准分

很多基于CTT的测验都用标准分来表示最后的测验结果，基本的标准分为z分数，z的取值范围为$(-\infty,+\infty)$，均数为0，单位标准差为1。通过线性转换，均数和标准差可以转换成所需要的形式。只要总体均数和标准差相等，两个测验分数就可以进行比较。标准分给人一个假象，就是测验分数为等距量表。很容易理解，计算标准分是基于对S的等距假设，因此，标准分的等距也就是一种假象。如果给S中的元素分配新的数据，随后的所有计算结果都将改变。

另一方面，标准分等距并不表明心理属性等距。举例说明，对于温度来讲，等于1℃的任何两个温度差的物理含义是一致的，而等于1的任意两个z分数差的心理含义并不相同，这是因为，与均数距离不同的z分数差，其分布曲线下的面积是不同的。

4.2　条目反应理论的Logit

Logit即对数胜算比，是条目反应理论中Rash模型最常用的结果表示方法，请参考www.rasch.org。对它的错误理解与标准分十分相似，此处不再赘述。

条目反应理论用极大似然法估计条目参数和潜在特质，极大似然估计所用的数据当然就是S，这本身就已经假定了S是等距量表。

5　小　　结

从测量学上讲，绝大多数心理测验只能达到顺序量表水平，不适合进行算术运算，因此，我们就武断地假定了心理测验达到了等距水平，并按等距量表进行各种处理。心理测验等距假设是指当用分数表示被试对条目的反应时，这些分数有相同的单位，相同的分数差表示了相同的距离。我们认为，CTT的标准分和条目反应理论的Logit只是等距假设的数学运算的结果，并没有消除心理测验的不等距性。我们还提出了条目质性和条目集群的概念，希望能够为心理测验研究提供新的思路。

参　考　文　献

1　Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill Book Company. 1956. 北星，译. 非参数统计. 北京：科学出版社，1986. 25-35

2　Anastasi A, Urbina S. Psychological testing. Prentice Hall Inc, 1997. 缪小春，竺培梁，译. 心理测验. 杭州：浙江教育出版社，2001. 5-13

3　Guilford JP, Fruchter B. Fundumental statistic s in psychology and education. 6th ed. New York: MeGraw-Hill, Inc, 1978

4　Kroenke K, et al. J Gen Intern Med, 2001, 16(9): 606-613

5　徐云，戴晓阳. 对《韦氏成人智力量表等几种心理测验修订中存在的问题》一文的商榷. 心理学报，1988，2：195-200

6　王文中，吳齊殷. 縱貫性研究中度量化的一些議題：以症狀檢核表SCL-90-R為例. 中華心理衛生學刊，2003，16(3)：1-30

7　刘仁刚，高北陵，等. Hiscock迫选数字记忆测验的修订和试用. 中国临床心理学杂志，2001，9(3)：173-175

8　刘仁刚. 条目反应理论简述. 中国临床心理学杂志，2009，17(1)：37-41，50

(收稿日期:2014-04-20)

(上接第844页)

10　汪向东，王希林，马弘. 心理卫生评定量表手册(增订版). 中国心理卫生杂志社，1999. 106-108，230-232，194-196

11　严虎，陈晋东，杨怡，等. 房树人测验在中学生自杀调查中的应用. 中国心理卫生杂志，2013，27(9)：650-654

12　王萍萍，许燕，王其峰. 汶川地震灾区与非灾区儿童动态房树人测验结果比较. 中国临床心理学杂志，2010，18(6)：720-722

13　严虎，陈晋东. 青少年图画心理分析手册. 长沙：中南大学出版社，2011. 42-49

14　张同延，张函诗. 揭开你人格的秘密：房、树、人绘图心理测验. 北京：中国文联出版公司，2007. 25-297

15　严虎，陈晋东. 青少年交往焦虑绘画特征研究. 神经疾病与精神卫生，2013，13(2)：184-186

(收稿日期:2014-03-12)