非参数项目反应理论在维度分析中的运用及评价

张 军
(北京语言大学 汉语进修学院,北京 100083)

摘 要:该文使用非参数项目反应理论的 Mokken 量表及其构建程序 MSP,探索性地分析 HSK[初中等]听力、语法结构和阅读三个部分中 40 个题的潜在维度,并籍此评价此方法的优劣。实验表明:题组是多维的,阅读题的区分能力和一致性最强,能有效地聚合成一类;听力题次之,语法结构题最差;此方法存在很多不足,尤其是题目区分能力对分类的干扰与界定分类阶段的标准问题。

关键词:非参数项目反应理论;潜在维度;MSP
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2010)03-0080-04

1 引言

非参数型项目反应理论(NIRT)自兴起、发展以来,多用于构建单维量表,其不同于逻辑斯蒂等参数型项目反应理论(PIRT)模型,不需要利用一个庞大的样本来做复杂的估算,但其介绍与应用在我国还比较少见。汉语水平考试(HSK)作为一种能力测验面临着构想效度的验证问题,但其构想效度的研究还不深入,尚未见到较多的证据,缺乏一个明晰的构想理论[1],方法又多为因素分析等传统方法,具有种种局限性。文章尝试使用 NIRT 来探索性地分析 HSK[初中等]的潜在维度结构,并籍此讨论 NIRT 作为维度分析工具的优劣。

2 维度研究的原理

过去十年里,用 NIRT 作为维度分析工具逐渐引起了研究者的兴趣。例如 Douglas,Kim,Roussos,Stout 和 Zhang[2] 研究了 Law School Admission Test 的维度;Scheirs 和 Sijtsma[3] 考察了 International Survey of Adult Crying 的维度性。

用 NIRT 研究测验维度的过程,实际上是在全部试题中,分析提取出若干个单维量表的过程,依据是量表适宜性系数(scalability coefficients)。具体细分为三种:试题 i 与试题 j 间的量表适宜系数 H_{ij};试题 H_i 与剩余试题全体间的量表适宜系数 H_i;试题全体的量表适宜系数 H。计算公式如下:

$$H_{ij}=\frac{Cov(X_i,X_j)}{Cov_{\max}(X_i,X_j)}=\frac{P_{ij}-P_iP_j}{P_i-P_iP_j}=1-\frac{P_i-P_{ij}}{P_i(1-P_j)}$$

$$H_i=\frac{Cov(X_i,R_{(i)})}{Cov_{\max}(X_i,R_{(i)})}=\frac{\sum_{j\neq i}(P_{ij}-P_iP_j)}{\sum_{j>i}(P_i-P_iP_j)+\sum_{j<i}(P_j-P_iP_j)}$$

$$H=\frac{\sum_i Cov(X_i,R_{(i)})}{\sum_i Cov_{\max}(X_i,R_{(i)})}=\frac{\sum_i\sum_{j\neq i}(P_{ij}-P_iP_j)}{\sum_i\sum_{i>i}(P_i-P_iP_j)+\sum_i\sum_{j<i}(P_j-P_iP_j)}$$

注:$R_{(i)}$ 指除 i 以外其他题的总分。

量表适宜性系数的一个优良性质是把保持高相关的,但又不同质的试题控制在低水平值上[4]。在非参数项目反应理论单调匀质模型中,用 H 系数分析抽取出的量表为 Mokken 量表。Mokken[4] 还提出仅当所有项目的 $H_i>c$ 时,那个量表才有用。c 是低限,可根据需要设定,至少为 0.3。当 $0.3\leqslant H<0.4$ 时,被认为是较弱程度的量表;当 $0.4\leqslant H<0.5$ 时,程度中等;当 $0.5\leqslant H$ 时,程度强。换言之,如果 H 在 0 到 0.3 之间,就不能相信项目组有足够共同的东西能将被试在一有意义的潜在特质上排序。

3 自动选题策略及 MSP

自动选题策略是构建 Mokken 量表的一种算法,而 MSP 就是根据其设计的操作程序。MSP 采用的算法基于项目对的 H_{ij} 系数,使用自底向上的序列选题策略,具体步骤如下:

第一步:在试题样本中选择 H_{ij} 大于 c、最高且检验显著的项目对。

第二步:计算每个备选试题与已选试题的 H_i 系数,选择与之系数最高且满足量表条件的试题。

第三步:重复第二步,直至无题可选。如果还有备选题目,那么开始从第一步重复,构建另一个单维量表,直至无题可构成另外的量表。

c 默认值为 0.3,H 系数显著性检验的 α 默认值为 0.05,为了减少偶然性风险,每步的检验方法使用 Bonferroni 校正检验。从本质看,这种序列选题

策略是一种顺序聚类算法，分析基础是 H 系数组成的矩阵[4,5]。H_{ij} 系数代表试题间的距离或相似性(distance/proximity)，代表项目间的关系强度；H_i 系数是已选题组和待选的某个题间的相似性度量，所以量表适宜性系数可看成是标准化的“相关系数”。

MSP 作为一个能保证区分度的选题策略，在构建量表时，为了保证试题的区分度，可能将测量了同维度、但区分度差的项目剔出去，这会造成分析维度时的两难困境：为了能保证项目正确归类，需提高 c 值；但提高 c 值后，又会将区分度差的，对同一维度敏感的项目剔出。

对此，Hemker 等[6]认为不存在一个适用于任何数据的独特 c 值，建议在执行 MSP 时，c 值从 0 开始，以 0.05 为步长，逐步增加到 0.55，这样就得到 12 种分类结果。如果试题是单维题组时，随着 c 值的增大，可能呈现出如下几个阶段：

1)绝大部分或全部项目归为一个量表；

2)形成一个较小的量表；

3)形成一个或几个小量表，同时许多项目被剔出。

如果是多维题组的话，则会表现为：

1)绝大部分或全部项目归为一个量表；

2)形成两个或多个量表；

3)形成两个或更多的小量表，同时剔出许多项目。

由此可见，在筛选过程的第一阶段单维题组和多维题组的表现相同，两者的分歧只是在第二、三阶段才显现出来，特别是第二阶段。这是由于第一阶段中 c 值过低，条件过于宽松造成的。虽然 Hemker 等的建议给了很大启发，却引出另一个值得商榷的地方：这三个阶段该如何定义和明确。这一问题目前尚无明确的理论研究和规定。不过究其本质，在三个阶段中 c 值的增大意味着量表强度的增加，所以根据 Mokken 提出的关于量表强弱程度的标准，采用如下标准：当 $c<0.3$ 时，为第一阶段；当 $0.3\leqslant c<0.5$ 时，量表从较弱到中等，为第二阶段；当 $0.5\leqslant c$ 时，为第三阶段。

4　测验维度的实验研究

根据非参数项目反应理论的原理，设计旨在探索性分析 HSK[初中等]部分试题潜在维度的实验。

4.1　实验材料

HSK[初中等]正式试卷的听力、语法结构和阅读部分试题，试卷代码为 M05N09X。其中听力题第一部分，从第 1 题到第 15 题；语法结构第一部分，从第 51 题到第 60 题；阅读第一部分中从第 81 题到第 95 题，共计 40 个题。每题均采用 0/1 计分，即答对为 1 分，答错为 0 分。

4.2　实验对象

从在复旦大学、南开大学、外企等 11 个国内考点，参加 2005 年 12 月 HSK(初、中等)考试的 12098 名被试中随机抽取 4924 人。为保证样本对总体的代表性，实验依据被试的试卷总分分层抽样。

被试总体的平均分为 100.8 分，标准差为 33.714，呈正态分布。抽取的被试样本，平均数为 98.8，标准差为 31.33，经 T 检验后样本平均数与总体平均数无显著差异，$p>0.05$。

4.3　实验方法

经分析，实验使用的 MSP 程序，以 0.25 为初始值，0.55 为终点值，0.05 为步长，逐步设定 c 的取值，对实验材料进行分析。

4.4　实验结果与分析

得出七种分类结果，具体统计量包含难度、观测 Guttman 误差、期望 Guttman 误差、H 系数等，详细可见附表($c=0.25$ 时，由于计算机配置的限制，只有分类结果，没有保存到各具体统计量)。为表达方便，将结果归纳如下。

表 1　分类结果表

C 值	
$C=0.25$	
量表 1	87 82 83 93 81 95 90 88 89 06 53 86 85 84 07 15 56 57 59 10 91
量表 2	09 03 01 12 13
量表 3	60 51
$C=0.3$	
量表 1	87 82 83 93 81 95 90 88 89 06 53 86 85 84
量表 2	15 10 03
量表 3	57 56
量表 4	09 01
$C=0.35$	
量表 1	87 82 83 93 81 95 90 88 89
量表 2	86 53 84
量表 3	15 10
量表 4	57 56
量表 5	06 04
$C=0.4$	
量表 1	87 82 83 93 81 95 90
量表 2	86 53
$C=0.45$	
量表 1	87 82 83
量表 2	95 90

量表 3 86 81

$C=0.5$

量表 1 87 82 83

$C=0.55$

量表 1 87 82

表 1 列出了 7 种分类结果，如前 4 行表示为：当 $c=0.25$ 时，选出三个子量表，每个子量表中包含若干试题。各量表及量表内试题的排列顺序是 MSP 按量表适宜性系数的大小筛选的顺序。其中“量表 1”等只是一个称谓，并不意味着所代表的实质相同。

40 个题的量表适应性系数都大于零，由程序的计算原理可推知，当 $c=0$ 时，40 个试题都会汇聚到一个单一的量表中。随着 c 值的增大，最开始形成的单一量表逐步分化成若干个子量表，如 $c=0.25$ 时，绝大部分试题归到第一类中，符合单维或多维题组第一阶段的特征描述。当 $0.3\leqslant c<0.5$ 时，量表从较弱到中等，MSP 始终都得到了两个或两个以上的小子量表。具体而言，$0.3\leqslant c<0.4$ 时形成了三个相对稳定的聚类中心：第 87 和 82 题、第 10 和 15 题、第 57 和 56 题。$0.4\leqslant c<0.5$ 时，量表成了中等强度，后两个聚类中心及相应的子量表消失，但量表 1 却继续分化成若干小量表，这符合多维题组第二阶段的特征描述。当 c 增大到 0.5 后，只留下一个小量表，直至只剩下第 87 和 82 题，这符合单维题组第三阶段的特征描述。虽然在第二阶段多维的特征很明显，但考虑到第三阶段的特征，很难判断题组是单维还是多维，因此还需要继续做更深入的分析。

在每种分类中，总有些试题不能进入任何子量表中，如当 $c=0.55$ 时，只有第 87 和 82 题组成的一个量表。纵观整个过程，进入量表的试题大部分是阅读题，少部分是听力题，语法结构题极少，如 $c=0.25$ 时，阅读题有 13 个，听力题有 8 个，语法结构题 5 个。在 c 达到 0.4 及以上时，量表中甚至只剩下阅读题。这说明 40 个题的区分力和一致性属阅读题最强，听力题次之，而语法结构题最差，这与 HSK 有关的题目分析研究结论相符。

并且每次聚类过程，都是第 87 和 82 题最先入选，说明这两个题的区分度和一致性极高。大多数阅读题依次聚合到以 87 和 82 为中心的分类中，而绝大多数听力和语法结构题与阅读题不在一个量表中。假设 87 和 82 题测量的是所谓的“阅读能力”，那么大部分听力题和语法结构题测量到的潜在特质是与“阅读能力”不同的特质。

听力题和语法结构题虽也有区分度和一致性较强的聚类中心，如听力题的 15 和 10 题（还有第 9 和第 1 题，第 6 和 4 题，但没有 15 和 10 题明显）；语法结构题的 57 和 56 题。他们在第二阶段量表在成为中等强度前还存在的，且相对独立，没有被同化到阅读题的聚类中。但随着 c 增加到 0.4 以上后，其再未出现过，只剩下阅读题的聚类。从整个分类过程的趋势看，这两类聚类中心所代表的潜在特质与阅读题代表的也不同。所以综合以上分析，有理由判断题组是多维的，而第三阶段之所以只留下一个小量表，最主要是因为听力和语法结构题的区分度和一致性非常差，无法“坚定地”保留下来，成为另一个小量表。

有的题在某种分类中没有出现，当改变 c 值，选择条件更苛刻时反而出现，并与原本在某一子量表中的试题重新构成另一子量表。比如第 4 题在 c 为 0.35 时，与原来在量表 1 的第六题重新组成量表 5。同理，有的题目本来与某些题在一个子量表，后来却脱离原来的量表，与别的试题构成区分力和一致性更高的新量表，如第 81 题。这是构拟最佳量表的过程，充分说明在探测题组维度时，没有一个绝对特定的 c 值，而要从小到大，逐步设定 c 的取值来综合判断，避免顺序选题策略中“能进不能出”的弊端。

与上一个现象大相径庭的是，随着 c 值的增大，有的题脱离原来的子量表后，并没有和别的题重新组合，而是从此“消失”了，如当 $c=0.35$ 时的第 88 和 89 题；当 $c=0.3$ 或 0.35 时的第 15、10、57 和 56 题，这很可能是因为虽然这些题目测量了相对应的潜在特质，但其区分度较小未能满足程序条件而被剔出的。

5 结论

5.1 从 M05N09X 号试卷中抽取的 15 个听力题、10 个语法结构题和 15 个阅读题构成的题组是多维的。在聚类过程中出现了具有代表性的三个聚类中心：第 87 和 82 题、第 15 和 10 题、第 57 和 56 题，其分别代表了三个不同的潜在特质维度。

5.2 阅读题的区分力和一致性程度最强，听力题次之，语法结构题最差。

5.3 听力、语法结构和阅读三个部分中，有的试题与本部分其他试题的一致性并不比与其他部分的试题的一致性高，测量到了与本部分其他试题测量到的特质所不同的特质，如第 53 题。

5.4 用顺序选题策略及其程序 MSP 侦测题组维度时，要从小到大，逐步设定 c 的取值，做综合判断，不存在唯一特定的 c 值。

5.5 区分力是影响试题归类的一个非常重要的因素，某个题即便测到了相同的特质，也能因为区分力小而不能聚合到同类中去。

6 研究展望

运用 NIRT 能侦测出部分试题对被试的区分能力及一致性程度，但也暴露出 NIRT 应用于维度分析的种种不足。

1）这种方法对维度的侦测目前还没有一个明显的指标或统计量，来明确试题的归类。只能通过 c 值的变换，大略分析分类的趋势，总结各阶段的特征，但如何划分各阶段，亦缺乏一个明晰的标准。

2）维度的分析不仅受到潜在特质的影响，而且试题的区分能力也是一个非常重要的因素，这是方法先天存在的限制性。

3）此方法旨在构建单维量表，试题间一定强度的共同变化（协方差）能用一个相同的潜在“因素”来解释，那么试题就测到了相同的特质。但如果不同的试题测量了彼此不同，却

存在高相关的潜在“因素”，也可能出现试题间非常一致的共同变化，以致误认为试题是单维的。这一情况在非参数反应理论中也很难分清。

4)在自动选题策略的选题方法上，通过 H 系数的高低选择试题，一旦选入就不能剔出。如果想重新尝试分类，唯一的方法就是增大 c 的取值，而这又往往受到区分力的影响，使程序面对区分力小的试题时失灵。

尽管非参数项目反应理论在分析维度时，目前仍存在很大的不足，但是仍可通过其对试题的一致性变化有个综合的了解，其在理论和选题方法上的不足还需要进一步去完善。

参考文献

1　张凯. 汉语水平考试结构效度初探. 见：《首届汉语考试国际学术讨论会论文选》编委会. 汉语考试国际学术讨论会论文选. 北京：北京语言学院出版社，1994.

2　Douglas J，Kim H R，Roussous L，et al.. LSAT Dimensionality analysis for December 1991，June 1992，and October 1992 Administrations Law School Admission Council Statistical Report，1999：95－05.

3　Scheirs J G M，Sijtsma K. The study of crying：Some methodological considerations and a comparison of methods for analyzing questionnaires. In：Vingerhoets A J J M，Cornelius R R. Eds. Adult crying. A biopsychosocial approach. Hove，England：Brunner－Routledge，2001.

4　Mokken R J. A theory and procedure of scale analysis. The Hague：Mouton/Berlin：De Gruyter，1971.

5　Loevinger J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs，1947：61，4.

6　Hemker B T，Sijtsma K，Molenaar I W. Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. Applied Psychological Measurement，1995：19.

The Usage of Non－parametric Item Response Theory on Analyzing the Latent Trait Dimensions

Zhang Jun

(The College of Advanced Chinese Training，Beijing Language and Culture University，Beijing 100083)

Abstract：In the thesis we try to analyze the latent trait(s)of 40 items in listening，reading and grammar structure parts of HSK (Chinese Proficiency Test)，by using Mokken scale and MSP in Non－parametric items response theory(NIRT)，in order to provide some evidence to construct validation of HSK，and evaluate this method. The research result shows that these items are multidimensional. The reading items can cluster efficiently in contrast to the listening and grammar structure items because of the discrimination and consistency，even through the latter two parts also have obvious clustering center. In addition，a lot of deficient exists in this method，especially the discrimination power of items，and the standard limiting the periods during the process of scaling items.

Key words：non－parametric item response theory；latent trait；MSP