

试题反应理论的介绍

攸关全国三十万国中生多元入学依据的九十年国民中学学生基本学力测验，除了依照标准化测验的编制程序、施测流程之外，还有在测验分数的计算与解释上，也融入现代测验(IRT)的精神与方法，这些无非是为了能从学生的答題结果以精确地估量学生的能力并能公平地作为升学的分发依据。

基本学力测验从各学科领域双向细目表的拟定，试题的设计与取样，到明确的施测程序、指导语、计分方式，乃至于试题事前测试(预试)以及信、效度的建立，完全依照标准化的测验编制程序来进行，以使测验能有最佳的测量效果。

加上从题库中抽取试题，组成正式测验进行考试，这种做法的优点是试题的相关讯息(包括难易度、鉴别力、所测量的能力指标...等)都已经事先知道了，因此可以依据每次测验的目标，挑选最佳的试题来进行测验，使测验能发挥最准确的测量效果。有关的现代测验理论与统计方法，大部分的社会人士都不太清楚，我们特别邀请心理计量学博士余民宁教授撰文为大家揭开现代测验(IRT)的神秘面纱。

余教授现任政大教育系教授兼任教师研习中心主任，其专长为多变量分析、教育研究法、教育测验与评量、测验编制，近十年来致力推动现代测验(IRT)的观念。藉由余教授一系列文章的介绍，从测验编制、题库建立、能力量尺与分数等化，都有详尽的解说，更能让您掌握测验理论的发展趋势以及明了IRT未来的应用，我们期待本专栏的设计，让大家对现代测验(IRT)有更正确的认识，对基本学力测验赋予更大的信任，更重要的是，企盼您给予我们更多的指教与期勉。

试题反应理论的介绍

- | | |
|----------------|--------------|
| 一、测验理论的发展趋势 | 九、测验分数的等化（上） |
| 二、基本概念和假设 | 十、测验分数的等化（下） |
| 三、试题反应模式及其特性 | 十一、题库的建立 |
| 四、能力与试题参数的估计 | 十二、计算机化适性测验 |
| 五、模式与数据间适合度的检定 | 十三、试题偏差的诊断 |
| 六、能力量尺 | 十四、精熟测验 |
| 七、讯息函数 | 十五、IRT的其他应用 |
| 八、测验编制 | 十六、IRT的未来 |

第一章

試題反應理論的介紹(一)測驗理論的發展趨勢....

政大教育系教授 余民宁 着

考试制度的创设虽然源自中国，绵延数千年后，世界各国争相采用，以作为建立

文官制度的选拔依据但是中国却一直没有针对「考试」这门学问进行比较科学化的量化分析，致使近代的心理计量学(psychometrics)却发展且发扬于外国，西风东渐后，才传入中国。

心理计量学是一门研究心理测验(psychological testing)与评断(assessment)的科学(Cohen, Montague, Nathanson, & Swerdlik, 1988, P.26)，是一门包括量化心理学(quantitative psychology)、个别差异(individual differences)、和心理测验理论(mental test theories)等研究范围的学问。比奈—赛门(Binet-Simon)的智力测验，可说是人类有史以来第一个心理测验，测验理论便是起源于此，并由此继续往前发扬光大，成为心理计量学的主要架构。

测验理论(test theory)（或全称叫「心理测验理论」）是一种解释测验数据间实证关系(empirical relationships)的有系统的理论学说，它的发展，迄今已迈入不同的新纪元，测验理论学者通常把它划分成二大学派：一为古典测验理论(classical test theory)——主要是以真实分数模式(true score model) (Gullikson, 1987; Lord & Novick, 1968)为骨干；另一为当代测验理论(modern test theory) ——主要是以试题反应理论(item response theory) (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hulin, Drasgow, & Parsons, 1983; Lord, 1980)为架构。这两派理论目前并行流通于测验学界，但试题反应理论却有后来居上，逐渐凌驾古典测验理论之上，甚至进而取而代之之势。

本文作者拟撰写一系列文章，介绍试题反应理论的主要理论内涵及其应用，在此之前，我们有必要从历史的观点，来回顾与展望测验理论的发展趋势，以明了测验理论发展的来龙去脉，这也正是本文的主要目的。

两派测验理论之比较

比奈—赛门的第一个心理测验问世后，正是心理计量学诞生之始，后经诸多学者（如：Cronbach, 1951; Guilford, 1954; Gullikson, 1987; Guttman, 1944; Lord & Novick, 1968; Richardson, 1936; Terman, 1916; Thurstone, 1929; Tucker, 1946）的研究与阐述，终于归纳形成古典测验理论等学说。

古典测验理论的内涵，主要是以真实分数模式（亦即，观察分数等于真实分数与误差分数之和，数学公式为 $X = T + E$ ）为理论架构，依据弱势假设(weak assumption)而来，其理论模式的发展已为时甚久，且发展得相当规模，所采用的计算公式简单明了、浅显易懂，适用于大多数的教育与心理测验数据，以及社会科学数据的分析，为目前测验学界使用与流通最广的理论依据。

然而，除上述各项优点外，古典测验理论却有下列诸项先天的缺失(Guion & Ironson, 1983; Wright, 1977)：

1. 古典测验理论所采用的指标，诸如：难度(difficulty)、鉴别度(discrimination)、和信度(reliability)等，都是一种样本依赖(sample dependent)的指标；也就是说，这些指标的获得会因接受测验的受试者样本的不同而不同，因此，同一份试卷很难获得一致的难度、鉴别度、或信度。
2. 古典测验理论以一个相同的测量标准误(standard error of measurement)，作为每位受试者的测量误差指标，这种作法并没有考虑受试者能力的个别差异，对高、低能力两极端组的受试者而言，这种指标极为不合理且不准确，致使理论假设的适当性受到怀疑。

3. 古典测验理论对于非复本(nonparallel)但功能相同的测验所测得的分数间，无法提供有意义的比较，有意义的比较仅局限于相同测验的前后测分数或复本测验分数之间。
4. 古典测验理论对信度的假设，是建立在复本(parallel forms)测量的概念假设上，但是这种假设往往不存在于实际测验情境里。道理很简单，因为不可能要求每位受试者接受同一份测验无数次，而仍然假设每次测量间都彼此独立不相关，况且，每一种测验并不一定同时都有制作复本，因此复本测量的理论假设是行不通的，从方法学逻辑观点而言，它的假设也是不合理的、矛盾的。
5. 古典测验理论忽视受试者的试题反应组型(item response pattern)，认为原始得分相同的受试者，其能力必定一样；其实不然，即使原始得分相同的受试者，其反应组型亦不见得会完全一致，因此，其能力估计值应该会有所不同。

一般说来，为了克服古典测验理论的缺失，才有当代测验理论的诞生。当代测验理论的内涵，主要是以试题反应理论为理论架构，依据强势假设(strong assumptions)而来，其理论的发展为时稍晚，理论模式也不断的在发展当中，所采用的计算公式复杂深奥、艰涩难懂，为一立论与假设均合理与严谨的学说，所适用的测验数据种类虽属有限，但深受测验学者的青睐，已有逐渐凌驾古典测验理论之上，甚至进而取而代之之势。

当代测验理论是为改进古典测验理论的缺失而来，它具有下列几项特点，这些特点正是古典测验理论所无法具备的(Hambleton, 1989; Hambleton & Cook, 1977; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980):

1. 当代测验理论所采用的试题参数(item parameters)（如：难度、鉴别度、猜测度等），是一种不受样本影响(sample-free)的指标；也就是说，这些参数的获得，不会因为所选出接受测验的受试者样本的不同而不同。
2. 当代测验理论能够针对每位受试者，提供个别差异的测量误差指标，而非单一相同的测量标准误，因此能够精确推估受试者的能力估计值。
3. 当代测验理论可经由适用的同构型试题组成的分测验，测量估计出受试者个人的能力，不受测验的影响(test-free)，并且对于不同受试者间的分数，亦可进行有意义的比较。
4. 当代测验理论提出以试题讯息量(item information)及试卷讯息量(test information)的概念，来作为评定某个试题或整份试卷的测量准确性，倒有取代古典测验理论的「信度」，作为评定试卷内部一致性指标之势。
5. 当代测验理论同时考虑受试者的反应组型与试题参数等特性，因此在估计个人能力时，除了能够提供一个较精确的估计值外，对于原始得分相同的受试者，也往往给予不同的能力估计值。
6. 当代测验理论所采用的适合度考驗值(statistic of goodness-of-fit)，可以提供考驗模式与数据间之适合度、受试者的反应是否为非寻常(unusual)等参考指标。

综合上述，当代测验理论似乎是绝对优于古典测验理论，但是事实上，当代测验理论被采用于解决真实测验数据者，比起古典测验理论广泛地被应用的情形而言，尚属少数，微不足道。其主要原因有下列诸点：

1. 当代测验理论系建立在理论假设严谨的数理统计学机率模式上，是一种复杂深奥、艰涩难懂的测验理论，这对于在数学方面训练有限的教育与心理学界学者而言，无非是一大挑战。阅读有关此理论之数学方面的研究报告与专书，已颇感困难，实在更难以深入将之发扬光大。
2. 多数当代测验理论学者都是出身自数学界或曾是数学主修者，或至少在数理统计学上训练有素者，他们偏爱对理论模式的探讨，远胜于对实际应用的推广工作。
3. 过去，计算机科技的进步有限，没有计算机软件包程序的实时配合，当代测验理论中对模式参数的估计，难以用手算或小型计算器顺利进行，因此，在应用上更受限制。
4. 有些古典测验理论的拥护者，对当代测验理论的研究与发展，所能获致之成效与应用性深表怀疑。为了证明与解释疑惑，当代测验理论学派的支持者，便更朝理论模式的量化技术方面探讨，致使当代测验理论的发展愈趋数学化、数量化、与计算机化。
5. 碍于严苛的基本假设，当代测验理论所能适用的教育与心理测验数据有限，并且需要大样本的配合，因此使得它的应用性大打折扣，未获一般测验使用者的全力拥护。

由上述两派测验理论的比较可知，古典测验理论虽然不够严谨，但理论浅显易懂，便于在实际测验情境（尤其是小规模数据）实施；当代测验理论虽然严谨，但理论艰深难懂，仅适用于大样本测验数据的分析。所以，这两派测验理论各有所长，在应用上也各有其限制，我们仅能静观测验理论的发展，逐步归纳出其未来的发展趋势。

测验理论的发展趋势

自从 Lord(1980)发表第一本以「试题反应理论」为名的专书后，当代测验理论正式以试题反应理论为其中心架构；在此之前，试题反应理论有个别称：「潜在特质理论」(*latent trait theory*)，由于潜在特质理论一词还包括「因素分析」(*factor analysis*)、「多元度量法」(*multidimensional scaling*)、与「潜在结构分析」(*latent structure analysis*)等，涵盖面甚广，无法精确反应出受试者在试题上的反应状况，因此，自 Lord 发表专书后，试题反应理论于是正式正名，且宣告诞生。所以自 1980 年后，测验学者逐渐以试题反应理论为当代测验理论的代表。

试题反应理论虽然自 1980 年才正式正名成立，然而在 30 和 40 年代，试题反应理论便已有初步的理论架构。其中，Tucker(1946)便是第一位使用「试题特征曲线」(*item characteristic curve*, 简称 ICC) 一词的心理计量学家，这一名词也逐渐成为试题反应理论的中心概念。兹将对试题反应理论发展有实际贡献的代表性作者及著作，条列简述于表一，由表一的内容便可获知试题反应理论的发展概况。

表一 对试题反应理论的发展有实际贡献的代表性作者和著作

作者（年代）	代表作及其贡献
Tucker(1946)	第一位提出试题特征曲线概念的人。
Lord(1952)	第一位导出两个参数常态肩形模式的参数估计公式，并考虑试题反应理论应用性的人。

Rasch(1960)	试题反应理论中 Rasch 模式的创始者，影响深远。
Lord & Novick(1968)	第一本介绍古典与当代测验理论模式的经典作品，引发学者对「潜在特质」概念的重视与研究。
Wright & Panchapakesan (1969) Samejima(1969)	美国地区第一篇介绍 Rasch 模式的参数估计法，并发展有名的 BICAL 计算机程序的代表作品。她的一系列作品描述新的试题反应模式及其应用，其中包含处理多分法与连续性数据的模式，甚至扩展到多向度的试题反应模式，为一艰涩难懂的重要著作。
Bock(1972)	提供许多估计模式参数的新概念。
Andersen(1973)	欧洲地区谈论测验模式的重要著作。
1976	Lord 等人创作第一版有名的计算机程序：LOGIST。
1977	Journal of Educational Measurement 第四季出版一册专门探讨试题反应理论的专辑。
Baker(1977)	第一篇评论试题反应模式参数估计法的文献探讨。
Wright & Stone(1979)	第一本描述各种 Rasch 模式理论及其应用的专书。
Lord(1980)	第一本以试题反应理论命名的专书，是当代测验理论发展的里程碑。
Weiss(1980)	第一本编辑成的论文辑，专谈试题反应理论的实际应用课题——计算机化适性测验。
Andersen(1980)	对测量模式参数估计法有贡献的方法学专论。
Bock & Aitkin(1981)	提出边缘的最大近似值估计法——EM 估计程序，对参数估计法的改进贡献不少。
Masters(1982)	第一位发表部份知识计分模式，对改进 Likert 式评定量表的计分与次序反应资料的计分贡献不小。
Wright & Masters(1982)	阐述 Rasch 模式的各种模式成员，证明皆与部份计分模式相通，对 Likert 式评定量表与次序反应数据的计分方式改进不少。
Mislevy & Bock(1982)	发表另一有名的计算机程序： BILOG。
1982	Applied Psychological Measurement 第四季出版一册专门探讨试题反应理论及其应用的进阶专辑。
Wainer & Messick(1983)	编辑而成的论文集，以表扬 Lord 一生对试题反应理论的贡献，并兼论该理论的应用与未来。
Weiss(1983)	编辑而成的论文集，专谈试题反应理论的应用与未来，并介绍它在计算机化适性测验上的应用。
Hambleton(1983)	编辑而成的论文集，专谈试题反应理论的模式与应用。
Hulin, Drasgow, &	为一本试题反应理论的教科书，增加对「适合度测

Parsons(1983)	量」概念的说明与应用。
Embreton(1985)	编辑而成的论文集，专谈试题反应理论的未来发展。
Baker(1985)	为一本导论性的试题反应理论教科书，专为没有数学训练基础的读者而作，并附有 CAI 的计算机教学磁盘。
Hambleton & Swaminathan(1985)	为一本进阶的试题反应理论教科书。
Crocker & Algina(1986)	谈论与比较古典与当代测验理论的导论性教科书。
Wainer & Braun(1988)	专谈有关效度方面的论文集，也谈试题反应理论在效度上的应用。
Linn(1989)	负责主编第三版的「教育测量」(Educational Measurement)，其中增加一章专门介绍并评论试题反应理论。
Freedle(1990)	专谈人工智能及其在当代测验理论上应用之论文集。
Suen(1990)	介绍各种测验理论方面的教科书。
Wainer 等人(1990)	专谈计算机化适性测验方面的入门书，也谈试题反应理论在计算机化适性测验上的应用。
Hambleton, Swaminathan, & Rogers(1991)	试题反应理论方面的入门书，适用于非数学主修的初学者阅读。

其实，随着近年来人类在计算机科技上的突飞猛进，各种适用于试题反应理论的计算机软件程序（如：目前最常用，也最有名的程序 BILOG 和 LOGIST 等）相继诞生与再版修订，已使得美国很多研究机构、地方政府机关、和私人团体，都率先采用试题反应理论作为他们编制测验、施测、计分、解释、与提供咨询服务的依据。

此外，表一所示的发展趋势可见，当代测验理论的发展趋势不外朝两个方向同步进行——理论的发展愈趋数学化与理论的应用愈依赖计算机。相信在可预期的将来，测验理论的使用者必须同时具备数学与计算机方面的良好训练，方能对试题反应理论的了解与应用驾轻就熟，而测验理论在愈趋专业化、专家化后，也唯有在专家或专家指导下方能推广应用试题反应理论，不过照目前的发展趋势来看，试题反应理论要取代古典测验理论是指日可待的事。

参考书目

1. Andersen, E. B. (1973). Conditional inference and models for measuring. Copenhagen: Mentalhygiejinsk Forlag.
2. Andersen, E. B. (1980) Discrete statistical models with social science applications. Amsterdam: North-Holland.
3. Baker, F.B. (1977). Advances in item analysis. Review of

- Educational Research, 47, 151-178.
4. Baker, F. B. (1985). The basics of item response theory. Portsmouth, NH: Heinemann.
 5. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
 6. Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443-459.
 7. Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). Psychological testing: An introduction to tests and measurement. Mountain View, CA: Mayfield.
 8. Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.
 9. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
 10. Embretson, S. E. (Ed.) (1985). Test design: Developments in psychology and psychometrics. Orlando, FL: Academic.
 11. Freedle, R. (Ed.) (1990). Artificial intelligence and the future of testing. Hillsdale, NJ: Lawrence Erlbaum Associates.
 12. Guilford, J. P. (1954). Psychometric methods. New York: McGraw-Hill.
 13. Guion, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. *Organizational Behavior and Human Performance*, 31, 54-87.
 14. Gullikson, H. (1987). Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates. (Originally published in 1950 by New York: John Wiley & Sons)
 15. Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
 16. Hambleton, R. K. (Ed.) (1983). Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.
 17. Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.
 18. Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.
 19. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.
 20. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.
 21. Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response

- theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.
22. Linn, R. L. (Ed.) (1989). Educational measurement (3rd ed.) New York: Macmillan.
23. Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.
24. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
25. Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
26. Mislevy, R. J., & Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Chicago: International Educational Services.
27. Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.
28. Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, 1, 33-49.
29. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.
30. Suen, H. K. (1990). Principles of test theories. Hillsdale, NJ: Lawrence Erlbaum Associates.
31. Terman, L. M. (1916). The measurement of intelligence. Boston, MA: Houghton Mifflin.
32. Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Bulletin*, 36, 222-241.
33. Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1-13.
34. Wainer, H., & Braun, H. I. (Ed.) (1988). Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.
35. Wainer, H. et al. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.
36. Wainer, H., & Messick, S. (Ed.) (1983). Principals of modern psychological measurement: A Festschrift for Frederic M. Lord. Hillsdale, NJ: Lawrence Erlbaum Associates.
37. Weiss, D. J. (Ed.) (1980). Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota.
38. Weiss, D. J. (Ed.) (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic.
39. Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological*

- Measurement, 29, 23-48.
40. Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA.

第二章

試題反應理論的介紹(二)

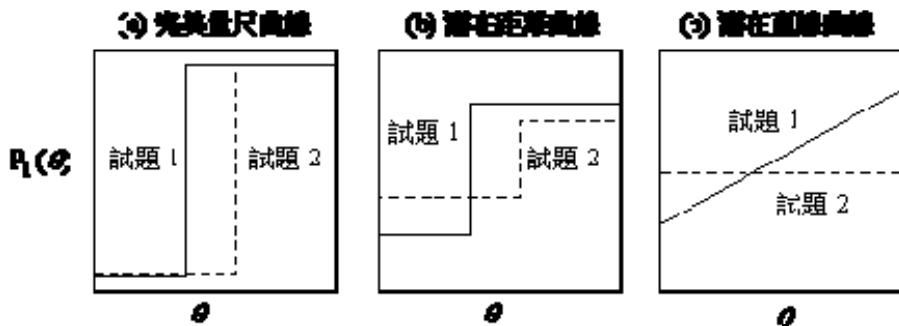
.... 基本概念和假設

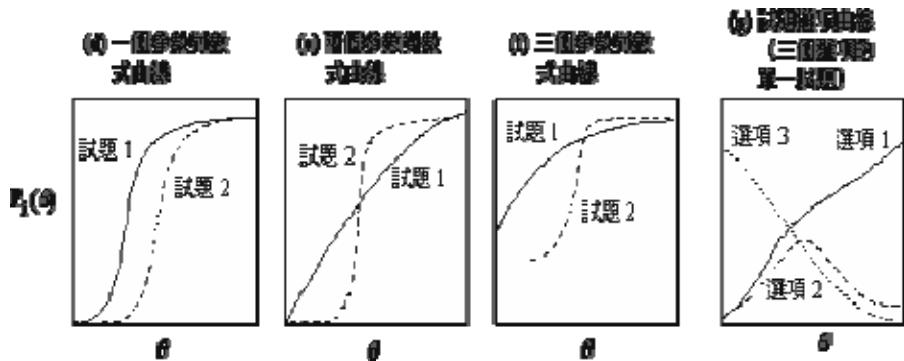
政大教育系教授 余民寧 着

基本概念

试题反应理论(item response theory)建立在两个基本概念上：(1)考生(examinee)在某一测验试题上的表现情形，可由一组因素来加以预测或解释，这组因素叫作潜在特质(latent traits)或能力(abilities)；(2)考生的表现情形与这组潜在特质间的关系，可透过一条连续性递增的函数来加以诠释，这个函数便叫作试题特征曲线(item characteristic curve, 简写为 ICC)。其实，我们把能力不同的考生得分点连接起来所构成的曲线，便是能力不同的考生在某一测验试题上的试题特征曲线，把各试题的试题特征曲线加总起来，便构成所谓的试卷特征曲线(test characteristic curve, 简写为 TCC)。因此，试题特征曲线即是一条试题得分对能力因素所作的回归线，这条回归线基本上是非直线的，但直线的试题特征曲线也是有可能的，端视所选用的试题反应模式(item response model)而定。

试题特征曲线所表示的涵义，即是某种潜在特质的程度与其在某一试题上正确反应的机率，二者之间的关系；这种潜在特质的程度愈高（或愈强），其在某一试题上的正确反应机率便愈大。在试题反应理论中，每一种试题反应模式就有其相对应的一条试题特征曲线，此一曲线通常包含一个或多个参数来描述试题的特性，以及一个或多个参数来描述考生的潜在特质；因此，所选用的试题反应模式所具有的参数个数及其数值的不同，所画出的试题特征曲线形状便不同。常见的曲线形状，如图一所示。





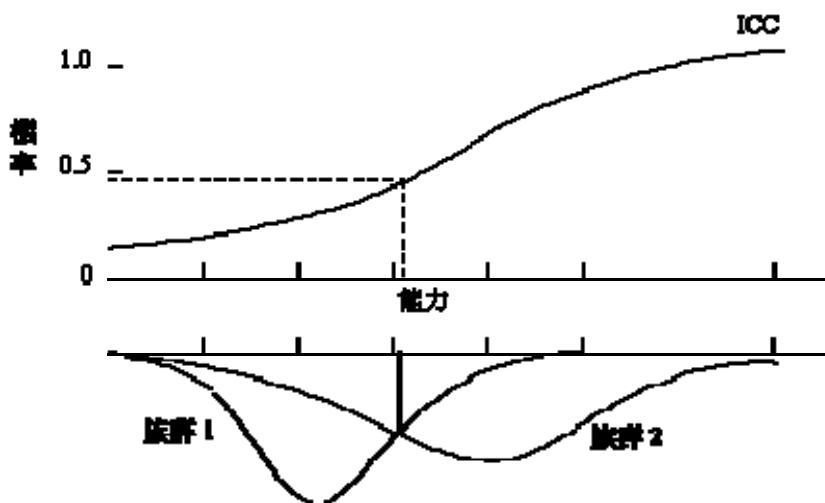
图一 七个不同的试题特征曲线例子（数据源：Hambleton & Cook, 1977）

图一所示的七种曲线中， θ 表示考生或受试者的能力或潜在特质， $P(\cdot)$ 则表示能力或潜在特质为 θ 的考生或受试者，其答对或正确反应某一试题的机率。例(a)所示，代表 Guttman 的完美量尺模式(perfect scale model)，它是一个阶段式函数(step functions)所形成的曲线，表示在某一关键值 θ^* 以右的机率为 1，以左的机率为 0；换句话说，这种模式具有完美的鉴别能力，而 θ^* 即为区别出有能力组和无能力组的关键值。例(b)为例(a)的一种变形，叫作潜在距离模式(latent distance model)，为社会心理学学者常用来测量人的态度的一种模式，其正确与不正确反应的机率，在 0 与 1 之间变动不已。例(c)所示即为古典测验理论下的试题特征曲线，截距的大小反映出试题的难度，而曲线的斜率即代表试题的鉴别度，在所考虑的条件相等的情况下，正确反应的机率与 θ 值成正比。例(d)到例(f)所示，即代表试题反应理论中的一个参数、两个参数、与三个参数的对数式模式(logistic model)，它们的涵义分别代表着：某一试题的正确反应机率除了受考生或受试者的 θ 值所决定外，并且分别受一个参数(即难度)、二个参数(即难度和鉴别度)、或三个参数(即难度、鉴别度、和猜测度)的试题参数所决定，其正确反应的机率值亦介于 0 与 1 之间。例(g)所示为特殊的试题反应模式，如：Samejima (1969) 的等级反应模式(graded response model)，Masters (1982) 和 Yu (1991) 的部份计分模式(partial credit model)等，即是采用试题选项特征曲线(item option characteristic curves)，所代表的意思是指试题中每一选项被选中的机率，它也是能力或潜在特质的一种函数，它有个基本假设，亦即就某一固定能力的考生而言，他 / 她在同一试题上所有的试题选项特征曲线的总和为 1。

试题反应模式不像古典的真实分数模式，它是可能作假的模式(falsifiable models)；换句话说，任何一种试题反应模式都有可能适用或不适用于某份特殊的测验数据，亦即模式可能会不当地预测或解释数据。因此，在应用试题反应理论时，我们必须先估计出模式与考生的参数值外，还需要考验模式与数据间的适合度(model-data fit)。这两者留待后文补充说明。

当某一种试题反应模式适用于某种测验数据时，一些试题反应理论的基本特性也会跟着产生。首先，从不同组的试题估计而得的考生能力估计值，除了测量误差外，

不会受所使用的测验种类的影响，亦即，它是试题独立(item-independent)的能力估计值；其次，从不同族群的考生估计而得的试题参数估计值，除了测量误差外，亦不受参与测验的考生族群的影响，亦即，它是样本独立(sample-independent)的试题参数估计值。上述两种特性，在试题反应理论中叫作「不变性」(invariant)，这些不变性是从把试题的讯息(information)考虑在能力估计的过程中，把考生能力的讯息考虑在试题参数估计的过程中而得。典型的试题参数不变性例子，如图二所示。在图二中，不管考生所来自的族群为何，只要他们具有相同的能力，他们答对（或正确反应）某一试题的机率便相同；由于某特定能力的考生答对某一试题的机率是由试题参数所决定，试题参数对这两族群的考生而言也必定相同。



图二 试题特征曲线与两族群考生的能力分配曲线

除了上述的特性外，试题反应理论还可以针对个别的（亦即每一位能力不同的考生或受试者）能力估计值，提供其测量的估计标准误(standard errors)，这点作法不同于古典测验理论仅提供所有考生单一的误差估计值的作法。此外，试题反应理论把能力测量的估计标准误之平方的倒数，定义为试题的讯息函数(item information function)，它可以用来作为评量能力估计值之精确度的指标，大有取代古典测验理论中「信度」(reliability)指标之势(Wright & Masters, 1982; Wright & Stone, 1979)。这些作法及概念续待后文补充说明。

基本假设

任何一条试题特征曲线所代表的涵义是：答对某一试题的机率，是由考生的能力和试题的特性所共同决定。因此，试题反应理论具有下列几项基本假设，唯有在这些假设都成立的前提下，试题反应模式才能被用来分析所有的测验数据。

(一)单向度(unidimensionality)：试题反应理论中的各种模式有个最常用的共同假设，那就是测验中的各个试题都测量到同一种共同的能力或潜在特质；这种单一能力或潜在特质（因素）必须包含在测验试题里的假设，便是单向度的假设。

其实，在实际的测验情境里，考生在测验上的表现情形很少是纯粹受到一种因素的影响，其他因素如：成就动机、考试焦虑、应试技巧、及人格特质等，也都会影响

到测验的结果；因此，试题反应理论中对测验必须具有单向度因素的基本看法，认为只要该测验具有能够影响测验结果的一个「主要成份或因素」(dominant component or factor)，便算符合单向度假设的基本要求，而这个主要因素所指的，即是该测验所测量到的单一能力或潜在特质。

适用于含有单一主要因素测验数据的试题反应模式，便称作单向度模式。适用于含有多种主要因素的试题反应模式，便叫作多向度(multidimensional)模式。多向度模式的数学公式复杂难懂，而且模式也还在发展中，本系列文章不拟介绍，有兴趣的读者可参阅 McDonald (1981) 和 Ackerman (1989) 的文章。

(二) 局部独立性(local independence)：它的涵义是说，当影响测验表现的能力被固定不变时，考生在任何一对试题上的反应，在统计学上而言是独立的；换句话说，在考虑考生的能力因素后，考生在不同试题上的反应间没有任何关系存在。简单地说，这意谓着涵盖在试题反应模式里的能力因素，才是唯一影响考生在测验试题上做反应的因素；这组能力因素代表整个潜在空间(complete latent space)，当单向度基本假设成立时，这整个潜在空间仅包含一种能力因素。

假设 θ 为能力因素， U_i 代表某位考生在第 i 试题上的反应， $P(U_i | \theta)$ 代表具有能力为 θ 的考生在第 i 试题上的反应机率，且 $P(U_i = 1 | \theta)$ 为正确反应的机率， $P(U_i = 0 | \theta)$ 为错误反应的机率，那么，局部独立性的涵义即是：

$$P(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta) P(U_2 | \theta) \cdots P(U_n | \theta)$$

$$= \prod_{i=1}^n P(U_i | \theta)$$

这条公式即是说明，对某一特定能力的考生而言，他 / 她在某份测验上的反应组型(response pattern)的机率，等于他 / 她在单独一题试题上反应机率的连乘积。例如，某位考生在一组三个试题测验的反应组型为 (1, 1, 0)，其中

$U_1 = 1, U_2 = 1, U_3 = 0$ ，那么，局部独立性所要表达的意思即为：

$$P(U_1 = 1, U_2 = 1, U_3 = 0 | \theta) = P_1(U_1 = 1 | \theta) P_2(U_2 = 1 | \theta) P_3(U_3 = 0 | \theta)$$

$$= P_1 P_2 Q_3$$

其中，

$$P_i = P(U_i = 1 | \theta) \text{ 且 } Q_i = 1 - P_i$$

由于 $P(U_i | \theta)$ 是一种条件机率(conditional probabilities)的表达方式，因此，局部独立性假设又称为条件独立性(conditional independence)假设。

这条公式即是说明，对某一特定能力的考生而言，他 / 她在某份测验上的反应组型(response pattern)的机率，等于他 / 她在单独一题试题上反应机率的连乘积。例如，某位考生在一组三个试题测验的反应组型为(1, 1, 0)，其中 ，那么，局部独立性所要表达的意思即为：

P

通常，当单向度假设获得成立时，局部独立性假设也会获得成立，就这一项涵义而言，这两个概念是相通的(Lord, 1980; Lord & Novick, 1968)，甚至于，即使数据不是单向度的，局部独立性也可以获得成立。只要整个潜在空间被界定清楚，亦即当所有影响表现的能力向度都考虑之后，局部独立性便可获得成立。局部独立性在下列情况下无法成立：影响测验表现的能力向度不只一种时，连锁性试题，以及试题本身提供作答的线索等，在这种情况下，试题反应模式也就无法适用于该笔测验数据。

(三)非速度测验：试题反应模式所适用的情况有个隐含的基本假设，那就是测验的实施不是在速度限制下完成的；换句话说，考生的考试成绩不理想，是由于能力不足所引起，而不是由于时间不够答完所有试题所致。由于这项假设是隐含在单向度假设里，所以不常被试题反应理论学者所提起，但是在选用试题反应模式时，这项基本假设亦必须要被考虑到才行。

(四)知道——正确假设(know--correct assumption)：如果考生知道某一试题的正确答案，他 / 她必然会答对该试题；换句话说，如果他 / 她答错某一试题，他 / 她必然不知道该试题的答案。当然，把正确答案填错在别的格子上以致整个试卷都错的例子，不在本假设所考虑的范围内，因为人为的疏忽不是任何测验理论所能顾及到的。此外，省略不答的试题(omitted items)和未答完的试题(unreached items)有所不同，前者是受能力影响所致，后者是受施测速度影响所致。本假设仅能适用于前者，它和前个假设一样，都隐含在单向度假设里，故殊少被提及。

参考文献

1. Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 13, 113-127.
2. Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.
3. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.
4. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.
5. McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.
6. Masters, G. N. (1982). A Rasch model for partial credit model. *Psychometrika*, 47, 149-174.

7. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.
8. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
9. Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
10. Yu, M. (1991). A two-parameter partial credit model. Unpublished doctoral dissertation of University of Illinois at Urbana-Champaign.

本文转载自研习信息 9 卷 (1 期), 5-9 页

[Back](#) [Home](#)

第三章

試題反應理論的介紹(三) 試題反應模式及其特性

政大教育系教授 余民宁 着

根据 Thissen & Steinberg (1986) 的分法, 所有的试题反应模式(item response models)依其基本假设与参数估计时的设限不同, 可以归纳为下列三大类:

1. 差异模式(difference models): 适用于次序反应的资料;
2. 除总模式(divide-by-total models): 适用于次序和名义反应的资料;
3. 左加模式(left-side added models): 适用于有猜题(guessing)可能的单选题反应数据。

虽然归类方式不尽相同, 到目前为止, 大多数已发展出来并且已在使用中的试题反应模式, 还是以适用于二元化计分(binary or dichotomous scoring)的性向或成就测验数据为主。本文的目的, 即在介绍试题反应理论中最常用的基本模式及其具有的特性。

基本的试题反应模式

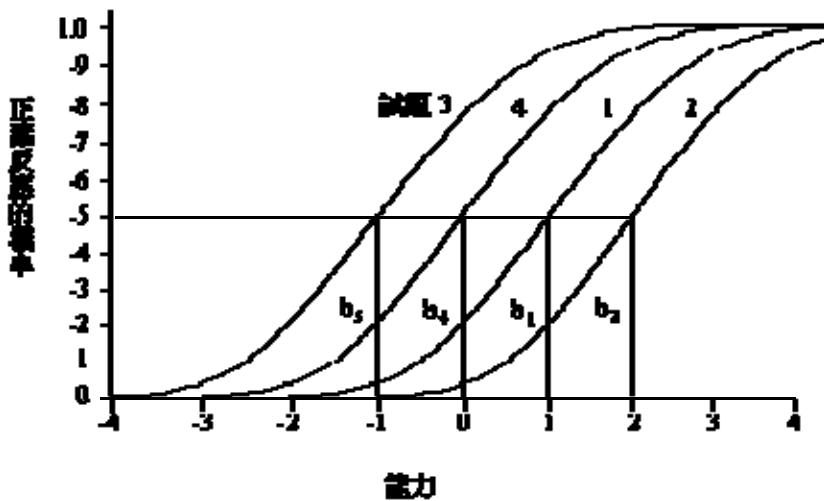
前文已经说过, 试题特征函数或试题特征曲线是用来描述测验所欲测量的潜在特质, 与其在试题上正确反应之机率间的一种数学关系; 因此, 每一种关系就有其相对应的一条试题特征曲线存在, 亦即是每一种试题反应模式都是用来描述特质与正确反应机率间的关系。常用的试题反应模式, 有下列三种, 每一种模式都依其采用的试题参数的数目多寡来命名, 都仅适用于二元化的反应数据(亦即, 正确反应者登录为 1,

错误反应者为 0 的数据)。

1. 一个参数对数形模式(one-parameter logistic model): 这个模式的数学公式如下所示:

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}} \quad i = 1, 2, \dots, n \quad (\text{公式一})$$

其中, $P_i(\theta)$ 表示任何一位能力为 θ 的考生答对试题 i 或在试题 i 上正确反应的机率; b_i 表示试题难度(difficulty)参数; n 是该测验的试题总数; e 代表以底为 2.718 的指数; 且 $P_i(\theta)$ 是一种 S 形曲线, 其值介于 0 与 1 之间。一个参数的试题特征曲线如图一所示。



图一 四条典型的一个参数试题特征曲线

根据公式一的定义, 试题难度参数 b_i 的位置正好座落在正确反应机率为 0.5 时能力力量尺(ability scale)上的点; 换言之, 当能力和试题难度相等时(即 $\theta - b_i = 0$), 考生答对某试题的机会只有百分之五十。当能力小于试题难度时(即 $\theta - b_i < 0$), 考生答对某试题的机会便低于百分之五十; 反之, 当能力大于试题难度时(即 $\theta - b_i > 0$), 考生答对某试题的机会便高于百分之五十。 b_i 值愈大, 考生要想有百分之五十答对某

试题的机会，他 / 她便需要有较高的能力才能办到，亦即该试题是属于较困难的题目。愈困难的试题，其试题特征曲线愈是座落在能力量尺的右方；反之，愈简单的试题，其试题特征曲线愈是座落在能力量尺的左方。图一所示，四条试题特征曲线的试题难度参数分别为 $a_1 = 1.0, a_2 = 2.5, a_3 = -1.0, a_4 = 0.0$ ，其值的大小，分别决定该四条曲线在能力量尺上的相对应位置，因此，试题难度参数有时又叫作位置参数(location parameter)。

理论上， b 值的大小介于 $\pm \infty$ 之间，但实际应用上，通常只取 ± 2 之间的范围。由图一所示， b 值愈大表示试题愈困难， b 值愈小表示试题愈简单。 b 值的概念符合常理的想法，但不同于古典测验理论中对难度 P 值的概念定义： P 值愈大表示试题愈简单， P 值愈小表示试题愈困难，其概念正好与常理的想法相反。这正是试题反应理论在解释试题特性上的一大优点。

由图一所示，四条曲线的形状是一致的，但在能力量尺上的位置各有不同，这点显示出：在一个参数模式下，影响考生在试题上表现好坏的试题特性只有一个，那就是该试题的难度。一个参数对数形模式并不把试题鉴别度(discrimination)指数考虑在内，其实，这种作法等于是假设所有试题的鉴别度都是相等的（通常设定为 1）。同时，它亦假设试题特征曲线的下限(lower asymptote)为零，亦即对于能力非常低的考生而言，他 / 她答对某试题的机会是零；换言之，一个参数对数形模式假设能力低的学生没有猜题猜中的可能，虽然考生们在单选题试题上往往会猜题。

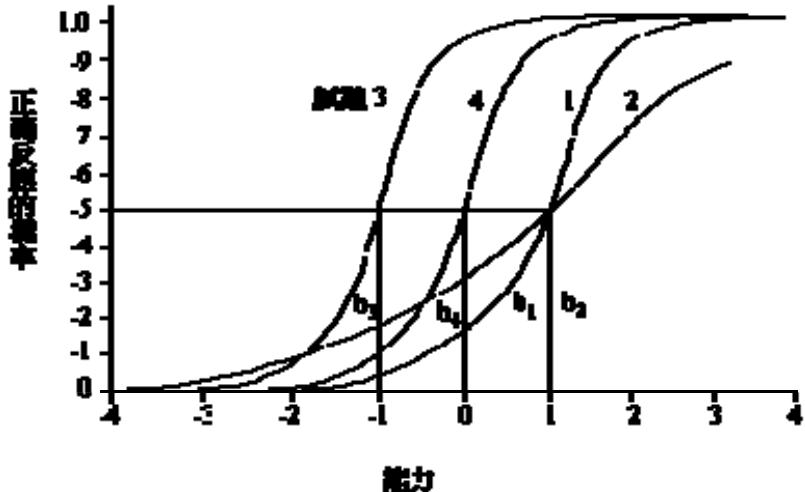
很明显的，一个参数模式的假设非常地严格。这些假设的适当与否，端视数据本身的特性和所欲应用该模式的重要性而定。例如，从一堆同构型颇高的题库(item bank)中选取相当容易的试题编制而成的测验，便非常符合这些假设的要求，这类情境常见于在有良好施测指导语下的效标参照测验(criterion-referenced tests)中。

一个参数对数形模式相通于 George Rasch (1960) 的模式，因此又有 Rasch 模式之称，以纪念这位丹麦的数学家在测验理论上所作的贡献。Rasch 模式通行于欧洲地区的心理计量学界，以及美国芝加哥大学等大学，有关 Rasch 模式的发展详情可参阅 Rasch (1960)、Wright & Stone (1979)、和 Wright & Masters (1982)。

2. 两个参数对数形模式(two-parameter logistic model)：这个模式的数学公式如下所示：

$$P_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}} \quad i = 1, 2, \dots, n \quad (\text{公式二})$$

其中，各符号的定义与公式一相同，唯多了一个参数：试题鉴别度(item discrimination) a_i ，它的涵义与在古典测验理论中的涵义相同，同是用来描述试题 i 所具有鉴别力大小的特性。典型的二个参数的试题特征曲线，可参见图二所示。



图二 四条典型的二个参数试题特征曲线

试题鉴别度参数 α 的值，刚好与在 b 点的试题特征曲线的斜率(slope)成某种比例。试题特征曲线愈陡(steeper)的试题比稍平滑的试题，具有较大的鉴别度参数值；换句话说，鉴别度愈大的试题，其区别出不同能力水平考生的功能愈好，亦即分辨的效果愈好。事实上，该试题能否区别出以能力水平为 θ ，上下两组（即高于 θ 和小于等于 θ ）不同能力考生的有效性，是与对应于 θ 量尺的试题特征曲线的斜率成某种比例。

理论上， α 值的范围在 $\pm\infty$ 之间，但学者们通常舍弃负的 α 值不用，因为该试题反向区别不同能力水平的考生，此外，带有负值 α 的试题特征曲线代表着：能力愈高的考生答对某试题的机率愈低，这似乎与学理相违背，所以负的 α 值不用。通常， α 值也不可能太大，常用的 α 值范围介于 0 与 2 之间； α 值愈大，代表试题特征曲线愈陡，试题愈有良好的分辨能力； α 值愈小，代表试题特征曲线愈平坦，正确反应的机率与能力间成一种缓慢增加的函数关系，亦即试题愈无法明显有效地分辨出考生的能力水平。

很明显的，二个参数对数形模式是由一个参数对数形模式延伸演变而来，亦即把试题鉴别度参数考虑进一个参数对数形模式里，便成为二个参数对数形模式。图二所示，四条试题特征曲线的试题参数分别为

$$\alpha_1 = 1.0, b_1 = 1.0; \alpha_2 = 0.5, b_2 = 1.0; \alpha_3 = 1.5, b_3 = -1.0; \alpha_4 = 1.2, b_4 = 0.0,$$

这些参数决定试题特征曲线的形状不会是平行的，因为有不同大小的试题鉴别度值存在的关系。当这四条试题特征曲线的 α 值都相等时，这些曲线便成平行的 S 形曲线，如图一所示；因此，我们可以这么说：一个参数对数形模式是二个参数对数形模式的一种特例，亦即把试题鉴别度参数都设定成一致时（通常设定 $\alpha_i = 1, i = 1, 2, \dots, n$ ），公式二的数学式子便简化成公式一的数学式子，这种说法于是成立。

由图二亦可知，这些曲线的下限值都是零，亦即二个参数对数形模式并不把考生的猜题因素考虑在内，这点假设与一个参数对数形模式雷同。猜题因素不存在的假设，往往使二个参数对数形模式适用于自由反应(free-response)的试题分析，或试题不太困难的单选题测验分析，对于有良好施测指导语的能力测验资料亦可适用。

二个参数对数形模式是由 Birnbaum (1968)修改自 Lord (1952)的原始二个参数常态肩形模式(normal ogive model)而来，由于它比常态肩形模式易于计算和解释，目前已取代常态肩形模式，而成为主要的试题反应模式。如果我们把公式二的分母与分子同时除以 $e^{-\theta_i}$ ，公式二也可以写成下列的横等式：

$$P_i(\theta) = [1 + e^{-\theta_i(\theta - \alpha_i)}]^{-1}$$

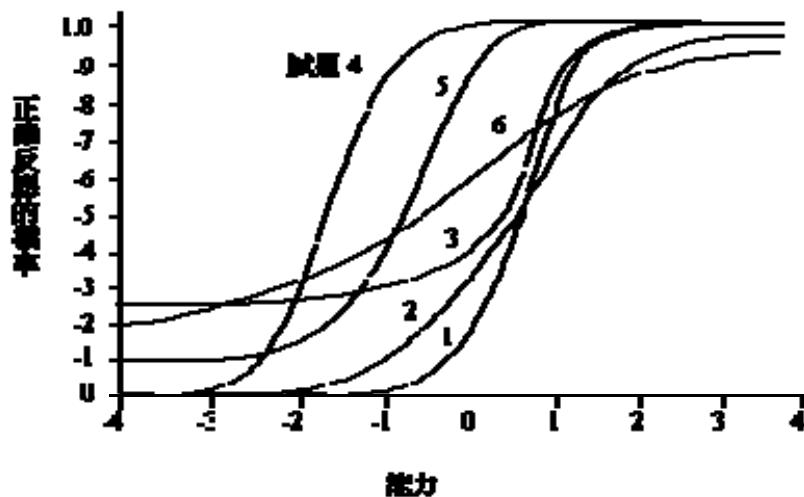
这个公式是二个参数对数形模式的另一种惯用表示方法。

3. 三个参数对数形模式(three-parameter logistic model): 这个模式的数学公式如下所示：

$$P_i(\theta) = C_i + (1 - C_i) \frac{e^{\theta_i(\theta - \alpha_i)}}{1 + e^{\theta_i(\theta - \alpha_i)}} \quad i = 1, 2, \dots, n \quad (\text{公式三})$$

其中，各符号的定义与公式二相同，唯多出一个参数：机运参数(pseudo-chance parameter) C_i 。这个参数提供试题特征曲线一个大于零的下限，它代表着能力很低的考生答对某试题的机率。

三个参数对数形模式是由二个参数对数形模式延伸演变而来，它多增加一个参数 C ，即是把低能力考生的表现好坏因素也考虑在模式里，当然，猜题可能是这些考生在某些测验试题（如：选择题）上唯一的表现行为。通常， C 参数的值比考生在完全随机猜测下猜答的机率值稍小，亦即 $C_i < 1/A_i$ ，代表试题 i 的选项数目。Lord (1974)认为，这是由于出题者通常会在试题中布置诱答选项的缘故，基于这项理由， C 不应该完全被视同「猜题参数」。三个参数的试题特征曲线如图三所示。



图三 六条典型的三个参数试题特征曲线

图三所示，六条试题特征曲线的试题参数分别为

$$a_1 = 1.8, b_1 = 1.0, c_1 = 0.0, a_2 = 0.8, b_2 = 1.0, c_2 = 0.0$$

$$, a_3 = 1.8, b_3 = 1.0, c_3 = 0.25, a_4 = 1.8,$$

$$b_4 = -1.5, c_4 = 0.0, a_5 = 1.2, b_5 = -0.5, c_5 = 0.1, a_6 = 0.4, b_6 = 0.5, c_6 = 0.15$$

这些参数决定这六条试题特征曲线的形状各不相同。其中，由第一条与第四条曲线的比较，可以显现出试题难度参数在试题特征曲线上的位置的重要性来：较困难的试题（如第1, 2, 3题）大多偏向能力量尺的高能力部份，而较简单的试题（如第4, 5, 6题）则多偏向能力量尺的低能力部份。由第1, 3, 4条与第2, 5, 6条曲线的比较，可以看出试题鉴别度参数对试题特征曲线的陡度(stEEPNESS)的影响力。最后，由第1条与第3条曲线的比较，**C**参数对试题特征曲线的形状也扮演着决定性的角色；同样的，试题3、5和6的下限的比较，也提供我们不少有关**C**参数的讯息。

其他常用的模式

除了上述三种基本的试题反应模式外，还有其他适用于非二元化数据的模式。例如：Bock (1972)的名义反应模式(nominal response model)是适用于名义反应数据的试题反应模式。Bock 的模式可用来分析单选题中每个选项被选中之机率；假设每个试题有 m 个选项，对每个 θ 而言，选择这 m 个选项之机率和为 1，这点是本模式的基本假设之一，另一个则是假设每个试题的 m 个选项间没有任何次序大小(ordering)的关系存在。当试题选项只有两个时，Bock 的模式便简化成二个参数对数形模式，所以 Bock 的模式是一种通用的模式(general model)。

另一类数据是多元化计分(polytomous scoring)的数据，一如 Bock 的模式所适用的数据，但数据本身多了一项特性：就是试题的选项（或反应）间具有次序大小的关系。适用于这类次序反应(ordered response)数据的模式有 Samejima (1969) 的等级反应模式(graded response model)，Andrich (1978a, 1978b, 1978c, 1978d, 1982)的二项式尝试模式(binomial trials model)和评定量表模式(rating scale model)，以及经 Masters(1982)归纳各种适用于次序反应数据的模式而提出的部份计分模式(partial credit model)，和经本文作者(Yu, 1991)扩充 Masters 的模式而成的「二个参数部份计分模式」。这类模式可用来作为分析李克氏量表(Likert scale)所属各种资料的工具，以改进社会科学研究的测量精确度，对社会及行为科学，甚至教育研究的量化方法学，具有着实的贡献潜能。

上述这些模式都是由基本的对数形模式延伸演变而来，由于新的模式还在层出不穷地诞生，本文无法一一详述，仅挑选基本的三种对数形模式作介绍，其余可参见 Thissen & Steinberg (1986)的分类说明。

参考书目

1. Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires. British Journal of Mathematical

- and Statistical Psychology, 31, 84-98.
2. Andrich, D. (1982). An extention of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, 47, 105-113.
 3. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories Psychometrika, 37, 29-51.
 4. Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.
 5. Rasch, G. (1980). Probability models for some intelligence and attainment tests. Chicago: The University of Chicago Press (Original edition publised in 1960).
 6. Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.
 7. Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.
 8. Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.
 9. Yu, M. (1991). A two-parameter partial credit model. Doctoral dissertation of University of Illinois at Urbana-Champaign (unpublished).

本文转载自研习信息 9 卷 (2 期), 6-10 页

第四章

試題反應理論的介紹(四)
....能力與試題參數的估計
(The Estimation of Ability and Item Parameters)

政大教育系教授 余民宁 着

应用试题反应理论的方法来分析某份测验数据的首要步骤，是估计我们所选用的试题反应模式的参数。有了满意的模式参数估计方法，整个试题反应理论的应用，才不致有滥用与误用等遗憾的情形发生。

前又说过，在试题反应模式里，正确反应的机率端赖两种因素，一为考生的能力参数，另一为试题参数。不论是能力或试题参数，二者都是未知的，我们唯一知道的是一群考生在一组测验试题上的作答情形（亦即是考生们的反应组型）。因此，参数估计的问题，便成为运用何种有效的方法，从现有的考生反应组型里，去推估适当的考生能力参数值和试题参数值的问题。这个问题很类似于回归分析中估计回归系数的问题，唯一不同者有两点：一为回归模式通常是直线的，而试题反应模式则是非直线的；另一为回归分析中的回归变项即（自变项）是观察得到的，而试题反应模式中的回

归变项（即 θ 变项）是观察不到的，需要进行估计才能得知。因此，假设 θ 为已知或观察得到的变项，则试题参数的估计问题，便相当于回归分析中去估计回归系数的问题；同样的，如果试题参数为已知，则能力参数的估计问题亦会变得相当地简单。本文的目的，即在讨论试题参数为已知下的能力估计，和能力参数为已知下的试题参数的估计方法：

能力参数的估计

假设某位考生在一份具有 n 个试题的测验上的反应组型(response pattern)为 $(U_1, U_2, \dots, U_j, \dots, U_n)$ ，其中 U_j 的值不是 1（代表正确反应），就是 0（代表不正确反应）。基于局部独立性的假设，上述观察到的反应组型的联合机率(joint probability)可以说是每一个试题反应机率的连乘积，亦即

$$P(U_1, U_2, \dots, U_j, \dots, U_n | \theta) = P(U_1 | \theta) P(U_2 | \theta) \cdots P(U_j | \theta) \cdots P(U_n | \theta)$$

或许也可以简化成

$$P(U_1, U_2, \dots, U_n | \theta) = \prod_{j=1}^n P(U_j | \theta)$$

由于 U_j 的值不是 1 就是 0，所以我们可以把近似值函数(likelihood function)表示成

$$P(U_1, U_2, \dots, U_n | \theta) = \prod_{j=1}^n P(U_j | \theta)^{\sigma_j} [1 - P(U_j | \theta)]^{1-\sigma_j}$$

或者简化成

$$P(U_1, U_2, \dots, U_n | \theta) = \prod_{j=1}^n P_j^{\sigma_j} Q_j^{1-\sigma_j} \quad (\text{公式一})$$

其中 $P_j = P(U_j | \theta), Q_j = 1 - P(U_j | \theta)$ 。

其实，公式一是某个反应组型的联合机率的表示公式；当这个反应组型为已知时，亦即 $U_j = u_j$ ，这种机率的解释方式便不再是恰当的，此时，对这种联合机率的表示公式便称作近似值函数，并且记作 $L(u_1, u_2, \dots, u_j, \dots, u_n | \theta)$ ，其中 u_j 代表在试题 j 上的实得反应。因此，

$$L(u_1, u_2, \dots, u_n | \theta) = \prod_{j=1}^n P_j^{u_j} Q_j^{1-u_j}$$

(公式二)

由于 P_j 和 Q_j 都是 θ 和试题参数的函数，近似值函数也是 θ 和试题参数的函数。

举例来说，假设我们有五位考生和五个试题，这些考生的反应组型和试题参数都是已知，详如表一所示。

表一 试题参数和五位考生在五个试题上的反应组型

试题	试题参数			考生的反应组型				
	a_i	b_i	c_i	1	2	3	4	5
1.	1.27	1.19	0.10	1	1	0	0	0
2.	1.34	0.59	0.15	1	0	0	1	0
3.	1.14	0.15	0.15	1	1	0	1	0
4.	1.00	-0.59	0.20	0	0	1	1	0
5.	0.67	-2.00	0.01	0	0	1	1	1

表一中的反应组型，1 代表答对该试题，0 代表答错该试题。以第三位考生为例， $u_1 = 0, u_2 = 0, u_3 = 0, u_4 = 1, u_5 = 1$ ，因此，这位考生的近似值函数可以公式二的表示方法表示如下：

$$L_3(u_1, u_2, u_3, u_4, u_5 | \theta) = (P_1^0 Q_1^0)(P_2^0 Q_2^0)(P_3^1 Q_3^1)(P_4^1 Q_4^0)(P_5^1 Q_5^0)$$

$$= Q_1 Q_2 Q_3 P_4 P_5$$

而第一位考生的近似值函数则可表示成：

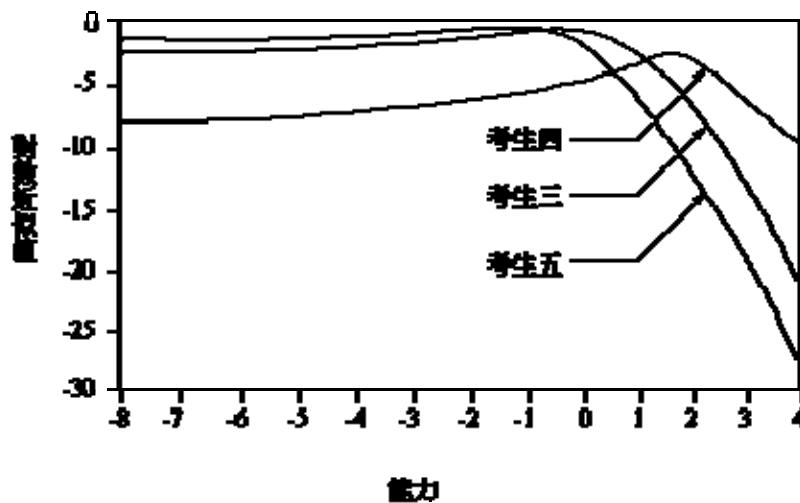
$$L_1(u_1, u_2, u_3, u_4, u_5 | \theta) = P_1 P_2 P_3 Q_4 Q_5$$

由于 P 和 Q 都是试题反应函数，它们的数学公式端视试题参数而定（在表一中的例子是三个参数对数形模式），而在本例中的试题参数已经是已知的情形，所以针对某个固定的 θ 值，便可算出其精确的近似值函数值；我们也可以根据不同的 θ 值，画

出其相对应的近似值函数图来。由于近似值函数是每个试题反应的机率之连乘积，而每个机率都是介于 0 与 1 之间，因此这个近似值函数的值会变得非常的小，不便于画图。有鉴于此，一个较好的量化方式，便是把近似值函数转换成自然对数的形式，再来进行估计参数或画图。因此，公式二取自然对数后(称作对数近似值 log-likelihood)可以写成：

$$\ln L(\theta | \alpha) = \sum_{j=1}^n [\alpha_j \ln P_j + (1 - \alpha_j) \ln(1 - P_j)] \quad (\text{公式三})$$

其中 α 代表试题反应的向量(vector)。根据考生能力及其相对应的对数近似值，可以图一来表示，其中第三位考生的对数近似值在 $\theta = -0.5$ 时最高，第四位考生在 $\theta = 1$ 时的对数近似值最高，而第五位考生的对数近似值在 $\theta = -1.5$ 时最高。此时，能够使某位考生的近似值函数(或相对应的对数近似值)达到最高点的 θ 值，便定义成该考生的 θ 的最大近似估计值 (maximum likelihood estimate, 简写成 MLE)。



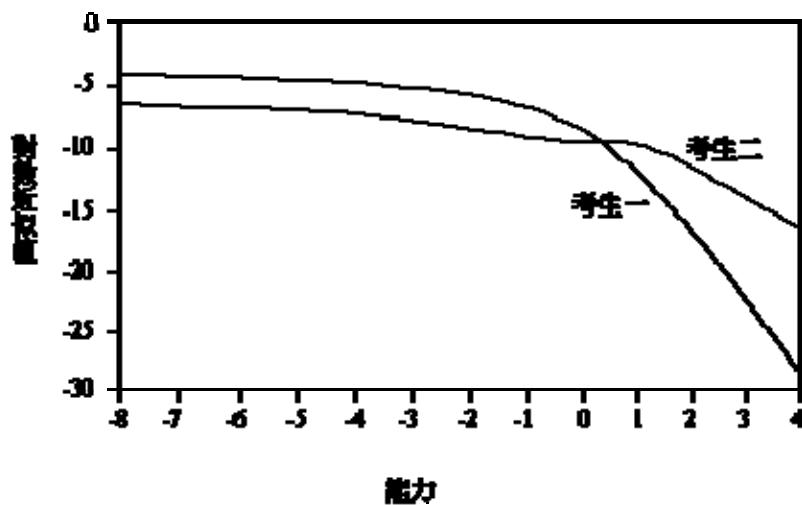
图一 三位考生的对数近似值函数

图一所示的图解法，并不是找出最大近似值函数的好方法，尤其是当考生人数和试题数目增多时，这种方法更是行不通。较有效的办法，便是利用近似值函数曲线的数学特性，亦即通过近似值函数最高点的函数斜率(以该曲线的第一阶导数来代表)必定为零。因此，我们可以利用微积分中求解函数的微分方式，把近似值(或对数近似值)函数方程式的第一个导数(first derivative)求出来，并且设定为零，再解这个方程式中相关参数的值，便可求得这些试题参数和能力参数的最大近似估计值。由于第一个导数方程式中往往同时包含一个以上的参数，因此，最大近似估计值无法由该方程式中直接求出，我们必须再求出其第二个导数，再套入 Newton-Raphson 的递归估计程序(iteration procedure)，透过现成的计算机程序(如：BILOG 或 LOGIST)，把参数的最大近似估计值求出，这整个估计过程可以详见 Hambleton & Swaminathan (1985, PP. 76-88)，本文不再赘述。

可惜的是，近似值(或对数近似值)函数有时不会出现只有一个固定的最大值，这种情况尤其在考生全答对或全答错试题时，便会产生。此时，能力参数的最大近似

估计值会变成 $\theta = +\infty$ 和 $\theta = -\infty$ 。以图二所示便可得知，第二位考生在 $\theta = 0.9$ 的点上的对数近似值最大，但其实在 $\theta = -\infty$ 的对数近似值才是真正最大的；同样的，第一位考生的最大对数近似值出现在 $\theta = -\infty$ 的点上。因此，这两位考生的最大近似估计值并不存在。

其实，出现上述这种现象的原因是由于这两位考生的反应组型都是特异的 (aberrant)：考生答对部份相当困难和有鉴别度的试题，却答错部份相当容易的试题。在这些情况下进行最大近似值估计法，最大的近似值往往无法收敛，以致无法获得一个明确固定的最大近似估计值。像这种特异的反应组型所产生的问题，通常只出现在三个参数模式上，而不会出现在一个和两个参数模式里 (Hambleton & Swaminathan, 1985)，有时也出现在 40 个试题以上的测验里。



图二 两个具有特异反应考生的对数近似值函数

最大近似估计值有个特殊的特性，那就是当它存在时，它具有大样本的渐近性 (asymptotic property)。由于我们所谈论的只是一位考生，渐近的意思是指逐渐增加的测验长度而言。当测验长度增加时， θ 的最大近似估计值，记作 $\hat{\theta}$ ，会呈现以 θ 为平均数的一种常态分配；这意谓着 $\hat{\theta}$ 的渐近分配会以真正的 θ 值为中心点，而呈现左右对称的常态分配，因此， $\hat{\theta}$ 值在较长的测验中是一种不偏的估计值 (unbiased estimate)。 $\hat{\theta}$ 的标准偏差，叫作标准误 (standard error)，记作 $SE(\hat{\theta})$ ，是 θ 的一种函数，表示成

$$SE(\hat{\theta}) = \sqrt{I(\theta)} \quad (\text{公式四})$$

其中的 $I(\theta)$ 叫作讯息函数 (information function)，我们将留待后文再介绍它的特性

及其对测验编制的重要性。由于我们无法事先知道 θ 值，所以我们必须将 $\hat{\theta}$ 值代入公式四里的 θ ，才能计算出 θ 所对应的讯息函数值。

有了 $\hat{\theta}$ 值等常态特性，我们也可以建立 θ 的信赖区间(confidence interval)。 θ 的 $(1 - \alpha)\%$ 的信赖区间，可以表示如下：

$$(\hat{\theta} - z_{\alpha/2} SE(\hat{\theta}), \hat{\theta} + z_{\alpha/2} SE(\hat{\theta}))$$

其中 $SE(\hat{\theta})$ 便是在 $\hat{\theta}$ 上的标准误，而 $Z_{\alpha/2}$ 是常态分配中上 $(1 - \alpha/2)$ 百分位数点；例如，95%的信赖区间的 $\alpha = .05$ ，而 $Z_{\alpha/2} = 1.96$ 。信赖区间可以提供研究者对 θ 估计值的精确性，一个参考的指标。

试题参数的估计

上一节所讨论的是假设试题参数为已知时，如何进行能力参数的估计。相反的，我们也可以假设能力参数为已知时，然后来行试题参数的估计。

假设每位考生的能力参数为已知，我们可以针对一群考生进行一组试题的施测，然后求出 M 位考生在每个试题的反应的近似值函数，即

$$L(u_1, u_2, \dots, u_M | \theta, a, b, c) = \prod_{i=1}^M P_i^{u_i} Q_i^{1-u_i} \quad (\text{公式五})$$

其中 a, b 和 c 是试题参数（假设以三个参数模式为例）。公式五是假设 M 位考生在每个试题上的反应是独立的，在这个假设满足后，公式五才算成立。

估计试题参数的方法与估计能力参数者雷同，仍然以常用的最大近似值估计法为之：我们分别针对 a, b 和 c 参数，求出近似值函数的第一阶导数，再把三个导数方程式设定为零，再同时解出这三个非直线方程式的解；对二个参数模式而言，有两个参数解，而一个参数模式则有一个参数解。接下来，可以 Newton-Raphson 的递归估计法，来求出这些方程式的解。当每位考生的能力参数为已知时，每个试题可以分别进行估计，而不必考虑其他试题的存在。所以，估计程序必须重复 M 次，每次估计一个试题。

其他估计方法与计算机程序

其实，在实际的估计情境中，我们往往无法事先得知能力和试题参数，因此，它们必须同时进行估计。我们可以采用上述的最大近似值估计法来进行参数的估计，这种同时进行估计能力与试题参数的最大近似值估计法，便叫作联合的最大近似值估计法(joint maximum likelihood estimation，简写成 JMLE)。由于详细的计算过程非常的繁琐，本文不拟在此讨论，有兴趣的读者可以参考 Hambleton &

Swaminathan (1985, 页 129-138)。

除了联合的最大近似值估计法外，尚有其他方法，如：边缘的最大近似值估计法 (marginal maximum likelihood estimation) (Bock & Aitkin, 1981)、条件化最大近似值估计法 (conditional maximum likelihood estimation) (Andersen, 1972; Rasch, 1960)、联合的和边缘的贝氏估计法 (Bayesian estimation) (Mislevy, 1986; Swamithan & Gifford, 1982, 1985, 1986)、启发式估计法 (heuristic estimation) (Urry, 1974)、和非直线因素分析法 (nonlinear factor analysis) (McDonald, 1967, 1989) 等，由于这些方法的数学公式艰深难懂，有兴趣的读者可以径行参阅该原始文献，本文不在此赘述。

估计能力和试题参数的过程虽然繁琐，但站在试题反应理论的应用观点来看，使用者不需要了解这些详实的估算过程，只要知道它们如何被估计出来，并且知道如何使用它们便可以。很值得庆幸的是，目前已有数种计算机程序问世，使用者只要会使用这些程序，便可获取能力与试题参数的估计值。有关这些计算机程序的简介，可参见附录一。

参考书目

1. Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.
2. Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, 11, 111-142.
3. Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Beston, MA: Kluwer.
4. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

附录一 目前常见的试题反应理论参数估计的计算机程序

程序名称	来源	适 用 模 式	估 计 方 法	所 需 计 算 机 配 备	优 点 (+) 缺 点 (-) 特 性 (★)
BICAL (BIGSCALE)	Wright 等 (1979); Wright 等 (1989)	一 个 参 数	非条 件 化最 大 近似 值	大 多 数 化最 大 近似 值 算 机	(+) 廉价的 (+) 提供标准 误 (+) 提供画图 / 适合度考 验 指 标
RASCAL	评量系统公司(1988)	一 个 参 数	非条 件 化最 大 近似 值	个 人 计 算 机	(+) 包括适合 度分析 (★) 并 入 Micro CAT 程 序包里
MICRO-	媒体交换科技公司	一 个	非条 件	个 人 计 算 机	(+) BICAL 的

SCALE	(1986)	参数 多 元 类 别	化最大 近似值	算 机	PC 版 (★)数据可以 放 入 电 子 表 格 里
ANCILLES	Urry(1974)	三 个 参 数	启 发 法	大 多 数 的 大 计 算 机	(+) 廉价的 (-) 常会删除 试 题 / 考 生 (-) 估计方法 不 严 谨 (★) 不 太 广 泛 使 用
ASCAL	评量系统公司(1988)	一 个 参 数 二 个 参 数 三 个 参 数	修改过 的 贝 氏 估 计 法	个 人 计 算 机	(+) 包 括 合 度 分 析 (+) 并 入 Micro CAT 程 序 包 里 (★) 使 用 贝 氏 估 计 法
LOGIST	Wingersky(1983); Wingersky 等 (1982)	一 个 参 数 二 个 参 数 三 个 参 数	非 条 件 化 最 大 近 似 值	IBM / CDC 大 计 算 机 (第 四 版)	(+) LOGIST 5 提 供 标 准 误 (+) 具 弹 性, 选 择 多 (+) 允 许 未 完 成 / 空 白 未 答 的 反 应 (-) 数据 输入 繁 琐 (-) 成 本 高 (-) 难 与 非 IBM 兼 容 设 备 联 线 (-) 设 定 许 多 限 制, 以 便 获 得 收 敛 的 参 数 估 计 值
BILOG	Mislevy & Bock(1984)	一 个 参 数 二 个 参 数 三 个 参 数	边 缘 的 最 大 近 似 值	IBM 大 计 算 机 个 人 计 算 机 版	(+) 选 择 性 贝 氏 估 计 值 (+) 可 避 免 极 值 的 估 计 值 出 现 (-) 在 大 计 算 机 上 的 执 行 成 本 很 高

					(一)错误的前置项假设会导致错误的估计值
NOHARM	Fraser & McDonald(1988)	一个参数 二个参数 三个参数	最小平方法	大多数的大计算机 个人计算机	(+)适用于多向度的模式 (+)包含残差值分析 (-)C 参数是固定的 (★)在美国地区的使用不广
MULTILOG	Thissen(1986)	多元类别		IBM 大型计算机	(★)把 BILOG 程序扩展成能够处理多元类别数据的程序
MIRTE	Carlson(1987)	一个参数 二个参数 三个参数	非条件化最大近似值	IBM 大型计算机 个人计算机	(+)适用于多向度的模式 (+)提供标准误 (+)包含残差值分析 (-)C 参数是固定的
RIDA	Glas(1990)	一个参数	条件化或边缘的最大近似值	个人计算机	(+)提供完整的考生与试题分析 (+)处理测验对换的不完整设计 (+)包含适合度分析

本文转载自研习信息 9 卷 (3 期), 6-12 页

第五章

試題反應理論的介紹(五) 模式與資料間適合度的檢定 (The assessment of model-data fit)

试题反应理论的特性与优点已在前几篇文章中介绍过了，这些特性与优点并不是随时都存在，它们只有在所选用的某种试题反应模式能够适用某种感兴趣的测验数据时，才能够存在；换句话说，在使用试题反应理论时，我们必须先检定模式与数据间是否具有满意的适合度(goodness-of-fit)，以确定所选用的模式能够适用于所分析的数据，方不致于误用或滥用试题反应理论的特性与优点。

检定模式是否能适用于所分析资料的方法有许多种，Hambleton & Swaminathan(1985)建议从下列三方面来作为判断的依据：

1. 模式对数据所具有的基本假设是否能够满足？
2. 模式所具有的特性（如：试题与能力参数的不变性）是否能如期获得？
3. 在使用真实和仿真数据下，模式预测力的正确性为何？

从上述三方面来进行模式与数据间的适合度检定，可以帮助试题反应理论的用户慎选适当的模式，作为应用试题反应理论的先前准备。以下便从上述三方面来介绍常用的检定方法。

模式假设的检定

我们可以根据不同的模式所具有的不同假设来进行检定。比较常见的检定假设和方法计有：

一、单向度假设的检定

1. 根据试题与试题间相关系数矩阵来进行因素分析，再依特征值(eigenvalues)大小，依序画出特征值分布图，再判断该图是否有一个明显的主要因素存在。
2. 比较真实测验数据与随机测验数据（样本数与试题数均相同）二者的试题间相关系数矩阵所画成的特征值分布图，如果单向度假设成立的话，则除了真实资料中的第一个特征值外，这两个特征值分布图应该会很相似，而真实数据的第一个特征值应该会比随机数据的第一个特征值还来得大。
3. 检查考生在能力量尺或测验分数量尺的不同范围内，其变异数——共变量矩阵或相关系数矩阵的局部独立性假设。当单向度假设（大约）成立时，该矩阵的非对角线元素值会很小，且趋近于零。
4. 针对试题间相关系数矩阵进行非直线的一个因素分析模式的因素分析，以检定它的残差值，并判断是否仍有其他因素存在。
5. 利用一种直接以试题反应理论为基础的因素分析的方法，来检定测验数据是否具有单向度的可能(Bock, Gibbons & Muraki, 1988)。
6. 检查某些看起来像是会违反假设的试题，看看它们是否表现出不同的功能。我们可以分测验的形式和总测验的形式，分别计算出这些试题的 b 值，如果单向度假设成立的话，这两种形式所计算出的 b 值所画成的图，应该会呈现直线分布的情形，并且具有可资比较的试题参数估计值的标准误。

除了上述的检定单向度假设所采用的方法外，Hattie(1985)曾提出八十八种指标，作为检定单向度假设的参考，他结论认为这些古老的心理计量学文献所提供的检定方法，多数都无法获得令人满意的结果，唯有以非直线的因素分析和残差值分析为基础的方法，才能获致最令人满意的检定结果。上述六种方法即是最具潜力的几种，其他可能的方法也正在发展中。

二、相等鉴别度指标假设的检定

这个假设检定通常仅适用于一个参数模式，因为它假设每个试题的鉴别度指标都相等。

我们可以从一种标准的试题分析中，逐题检视试题与测验分数间相关系数（二系列相关或点二系列相关系数）的分配，如果每个分配都呈现同质形状时，我们所选用的模式便算符合相等的试题鉴别度假设。

三、最小猜测度假设的检定

这个假设检定通常也只适用于一个和二个参数模式，因为它们均假设猜测度的可能性是微乎其微，甚至于完全没有。检定的方法至少有下列三种：

1. 我们可以检查低能力组考生在最困难试题上的表现情形，如果他们的表现水平是趋近于零，则这个假设可算是获得满足。
2. 我们也可以诉诸试题与测验分数间的回归线图的帮助。测验得分低的考生若倾向有接近于零的表现水平，则这个假设亦算是获得满足。
3. 我们也可以检视测验难度、时间限制、与试题的编排格式等，以检定猜测对测验表现的可能影响力。

四、非速度（难度）测验假设的检定

这个假设和单向度假设一样，均适用于所有的试题反应模式。我们可用至少下列三种方法之一来加以检定：

1. 我们可以比较没有回答的试题数之变异数和答错的试题数之变异数，当这个假设满足时，这项比值应该是接近于零。
2. 我们也可以比较在有时间限制下和没有时间限制下的考生测验分数，如果这两次考试的表现情形具有高度的重迭部份，则表示这个假设获得满足。
3. 我们也可以比较答完全部试题的考生百分比、答完百分之七十五试题的考生百分比、和被百分之八十考生答完的试题数，当几乎所有的考生答完几乎所有的试题时，速度便可被判定为不是影响测验表现的一个重要因素。

模式特性的检定

最常检定的两种模式参数的特性为：能力参数的不变性和试题参数的不变性。

一、能力参数估计值的不变性之检定

我们可以拿不同测验试题样本所得的能力估计值来作比较（例如：比较困难与简单的试题，或由题库中抽取不同内容范围所组成的测验所估计出的能力参数估计值）。

如果该估计值所相对应的测量误差间差异不大时，不变性的特性便算是符合。

二、试题参数估计值的不变性之检定

我们可以比较两组或多组受试者（例如：男人和女人；黑人、白人、和西班牙裔人；教学组别；高分与低分的考生；不同地区来应考的考生等）接受某种测验后，所获得该测验的试题参数估计值（例如： b 值、 a 值、或 c 值）。根据两组参数所画成的分布图，除了因样本大小所造成的分散误差外，这图应该是呈直线分布，且基线是由两个随机的相等样本所建立，若此，则参数估计值的不变性才算存在。

总之，两组模式参数（即根据同一批受试者在两种测验试题上的反应资料，所求得之能力参数估计值，和同一批测验试题让两组受试者施测后，所求得之试题参数估计值）所画成的分布图，可用来判断该分布图是否呈直线分布情形，若呈近似斜率为 1，截距为 0 的直线，则可说是某个试题反应模式适用于该份测验数据，且具有模式参数不变性等特性。

模式预测力的检定

另一种检定模式与数据间适合度的作法，便是进行试题残差值的分析。我们可以挑选一个合用的试题反应模式，并且估计出试题与能力参数，和求出各种不同能力组考生的表现情形，接着就可以比较预测的结果和真实的结果(Kingston & Dorans, 1985)。

某组考生在实得的试题表现(observed item performance)与期望的试题表现(expected item performance)之间的差距，便叫作原始残差值(raw residual)，记

作 r_i^* 。其数学公式如下：

$$r_i^* = P_{ij} - E(P_{ij}) \quad (\text{公式一})$$

其中， i 代表试题， j 代表某组考生的能力组别， P_{ij} 便是第 j 个能力组别在第 i 个试题上正确反应的实得百分比，而 $E(P_{ij})$ 则是在所选定（假设）的试题反应模式下正确反应的期望百分比。我们可以估计出假设的模式参数估计值，再利用这些估计值去计算一个正确反应的机率，这个机率便用来作为某个能力组别的正确反应的期望百分比。

使用原始残差值有个缺失，那就是无法顾及某个能力组别内期望的百分比正确分数的抽样误差。为了顾及这项误差，我们可以将原始残差值除以期望的百分比正确分数的标准误，以将原始残差值转换成标准化残差值(standardized residual) z_i^* 如下：

$$z_i^* = \frac{P_{ij} - E(P_{ij})}{\sqrt{E(P_{ij})(1 - E(P_{ij})) / N_j}} \quad (\text{公式二})$$

其中 N_j 是在能力组别为 j 的考生人数。

当我们选择试题反应模式时，原始残差值、标准化残差值、或两者的分析，可以提供许多参考的讯息。下列便是检定模式预测力常用的方法：

1. 检查模式与数据间适合度之残差值和标准化残差值。决定模式是否具有适合度，有助于挑选一个令人满意的试题反应模式(Ludlow, 1985, 1986)。
2. 在假设所有的模式参数估计值都正确的前提下，我们可以比较实得的与期望的测验分数的分配，卡方统计数（或其他统计数）或图解法可用来呈现这种比较结果。
3. 我们也可以检验试题替换的影响、练习的影响、测验的速限和作弊的影响、疲劳、课程、模式选择不当、指导语的前后时效、认知处理的变项、以及其他会违害试题反理论结果效度的不良影响，并且利用这些证据作为挑选某种适当的试题反应模式的参考。
4. 画出能力估计值和其相对应的测验分数间的分布图。当适合度的指标落在可被接受的范围内时，除了少数的数据点在测验特征曲线（反映出测量误差）周围作零星分散外，该数据分布图应该呈现出强烈的直线关系才对。
5. 运用许多统计考验的方法来检定整个模式、试题、或个别受试者的适合度。
6. 使用计算机仿真的方法来比较真实的与估计的试题与能力参数。
7. 利用计算机仿真的方法来检定模式的韧性(robustness)，例如，我们可以研究单向度的试题反应模式能否适用于多向度的资料(Ansley & Forsyth, 1985; Drasgow & Parsons, 1983)。

实际说来，为了计算残差值，我们通常把能力量尺分割成等距的段落 10 至 15 个区间。这些区间必须要够宽，以免落在这区间的考生数过少，因为样本数过少所得的统计数会不稳定；同时，这些区间也必须够窄，才能使得落在这区间的考生在能力上是属于同性质的。

接下来是计算实得的百分比正确分数：算一算在某一能力组别内的考生答对某试题的总数，再除以该能力组别内的考生总人数。同时，习惯上是以每一能力组别的组中点来代表该组别的能力参数值，然后以该值来计算某个正确反应的机率，并求出每一能力组别中每一位考生在某个正确反应上的机率，这些机率的平均值即当作该能力组别的期望百分比。有了实得的百分比正确分数和期望的百分比正确分数之后，我们便可代入公式二进行某种统计考验。

常用的统计考验方法是卡方考验(chi-square test)。Yen(1981)提出的 Ω_i 指标，便是一种典型的卡方考验所用的统计数指标，它可以用来检定模式是否适合数据。某个试题 i 的 Ω_i 统计数为：

$$\Omega_i = \sum_{j=1}^n \frac{N_j [P_j - E(P_j)]^2}{E(P_j)[1 - E(P_j)]} \quad (\text{公式三})$$

$$-\sum_{j=1}^m z_j^2$$

其中，根据能力估计值的不同，考生共可分成 m 个能力组别。 χ^2 统计数将成为一种以 $m - k$ 为自由度的卡方分配，其中的 k 即是试题反应模式的参数个数。如果所计算出的 χ^2 值大于所查表的临界值，我们便可以推翻试题特征曲线（或试题反应模式）适合数据的虚无假设，而该建议寻找另一个较佳的模式才对。

总之，一个检定模式与数据间适合度的方法，最理想的是包含(a)设计和执行各种分析，以便检查不适合情况的可能型态，(b)仔细考虑通盘的结果，(c)根据所欲应用的范围，判断模式是否合适。而分析的过程应包括对模式的假设、模式的特性、和模式预测力与实际数据间的差异等之检定，之后，再用统计考验的方法来检定虚无假设是否成立，以便提供统计讯息，作为挑选一个适当模式的参考。

参考书目

1. Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
2. Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.
3. Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
4. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.
5. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.
6. Hattie, J. A. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139-164.
7. Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement*, 9, 281-288.
8. Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement*, 45, 851-859.
9. Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10, 217-229.
10. Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

第六章

試題反應理論的介紹(六)能力量尺 (The ability scale)

政大教育系教授 余民宁 着

测验的最终目的往往在于给考生打分数，该分数便代表考生习得（或熟习）某种技能的程度。因此，谨慎地打分数以及小心翼翼地解释测验分数，便成为教育或心理测量中一个重要的课题。

在古典测验理论里，考生在某份测验上答对（或正确反应）的题数和，即代表他在该测验上的真实分数(true score)的不偏估计值(unbiased estimate)，亦即是他在该测验所测量之技能上的表现程度。而在试题反应理论里，考生的能力并不是由答对的题数和来表示，而是必须估计的；亦即经由某种适当模式以及考生的反应组型(response pattern)，来估计出考生应有的能力估计值。

一般说来，决定考生能力估计值大小的常用步骤，如下所述：

1. 先取得考生在一组试题上的反应资料，答对者给 1 分，答错者给 0 分。
2. 若试题参数（如：难度、鉴别度、及猜测度）为已知时，可用前文所提到的最大近似值估计法(maximum likelihood estimation)来进行估计考生的能力。
3. 若试题参数为未知时，则试题与能力参数就必须一起同时进行估计，此时亦可用最大近似值估计法来估计。
4. 再将估计好的能力估计值，经由直线或非直线转换，以便换算成较为一般大众所熟悉的量尺分数，增加解释测验分数的便利；例如：TOEFL 分数便是一例。

经由上述步骤，我们可以获得考生的能力分数，有助于我们解释考生在某份测验上的表现好坏。然而，在考虑解释能力分数的意义之前，能力分数的本质是什么？什么样的分数转换才有意义？该用何种量尺来表示？我们也必须要有所了解才行，方不致误用或滥用了试题反应理论的优点。

能力量尺的本质

前面说过，答对题数和分数 X 是真实分数 τ 的不偏估计值，亦即是

$$E(X) = \tau \quad (\text{公式一})$$

答对题数和分数 X 除以总题数（亦即经过直线转换），即可获得一个比例正确分数 (proportion-correct score)。当某测验包含许多分测验，且每一个分测验都包含不同的题数，测量到不同的目标时，使用比例正确分数则较具有意义与适当性。这种情形通常出现在效标参照测验(criterion-referenced test)里，而在常模参照测验(norm-referenced test)里，往往需要使用别种直线转换，才能获得所需的标准分数。当然，答对题数和分数 X 也可以透过非直线的转换，以换算成标准九分或百分位数等分数，以便于进行比较考生在某测验上的表现差异情形。

然而，分数 X 有个天生的缺点，那就是它不是一种试题独立的分数，因此，经过转换过的分数也不是一种考生族群独立的分数，亦即它会受到不同试题与不同考生的反应的影响。另一方面，能力分数 θ 却具有试题独立与样本独立等不变性的特质（请参考前文关于试题反应理论的基本概念之说明），它与 X 分数不同。所以，我们可以使用 θ 分数来比较回答不同试题的考生能力，而 θ 分数所用的量尺(scale)也可以被看成是测量特质或能力的绝对量尺。

其实，测验试题所欲测量的特质或能力，可以被广义的定义成态度或成就、一种狭义的成就变项（如：四则运算的能力）、或一种人格变项（如：自我概念、成就动机等），它们未必是天生的、或是一成不变的。事实上，能力或特质一词被看成是考生的一种固定特征时，多少都有一点不恰当或误导的意味在里头。在许多情境里，使用精熟程度(proficiency level)一词，也许会比较恰当些。

另者，定义 θ 分数所用的量尺，又具有什么样的本质呢？很明显的，观察分数 X 不是定义在比率量尺(ratio scale)上，也不是定义在等距量尺(interval scale)上，充其量，它最多仅被定义在次序量尺(ordinal scale)上。同样的， θ 分数亦被定义在次序量尺上。然而在某些情境中， θ 量尺被作为有限的比率量尺的解释，也是有可能的。

θ 量尺的转换

转换可分成两种：直线转换和非直线转换。转换的目的在于使测验分数的解释和涵义的了解，能广被一般大众所接纳。以下就以这两种转换来说明 θ 量尺的涵义。

读者们可还记得：在试题反应理论里，正确反应的机率是以试题反应函数 $P(\theta)$ 来表示。以二个参数模式为例：

$$P(\theta) = e^{(\alpha\theta + \beta)} / [1 + e^{(\alpha\theta + \beta)}] \quad (\text{公式二})$$

若将 θ 、 b 、及 a 加以转换成： $\theta' = a\theta + b$, $b' = ab + b$, 和

$a' = a/a$, 则

$$P(\theta') = P(\theta) \quad (\text{公式三})$$

亦即经过直线转换后，一个正确反应的机率不会改变；它意谓着，只要试题参数也经过适当的转换，我们便可针对 θ 量尺进行直线转换，而仍不改变其正确反应的机

率值。

例如，Woodcock(1978)的心理教育测验库(Woodcock-Johnson Psycho-Educational Battery)所用的量尺，便是以一个参数模式求得之 θ 值，经转换成以9为底的对数量尺：

$$W_s = 20 \log_9(\sigma^\theta) + 500 \quad (\text{公式三})$$

或

$$W_s = 9.1\theta + 500 \quad (\text{公式四})$$

因此，它是一种直线量尺(linear scale)。同理，试题难度也可以转换成

$$W_s = 9.1\theta + 500 \quad (\text{公式五})$$

而 W_s 量尺的特性之一便是 $(W_s - W_0) = 20, 10, 0, -10, -20$ 的差值，其正确反应的机率分别刚好是.90, .75, .50, .25 和.10。Wright(1977)曾把这个量尺修改成

$$W = 9.1\theta + 100 \quad (\text{公式六})$$

并把它叫作「智慧」量尺(WITs scale)。

有时候在某些情境里，进行非直线转换(nonlinear transformation)反而有助于我们对参数的估算和解释。兹以一个参数模式为例，说明如下：

$$P(\theta) = e^{(\theta-\delta)} / [1 + e^{(\theta-\delta)}] \quad (\text{公式七})$$

如果我们把 θ 和 b 值经由非直线方式转换成新的 θ' 和 b' 值如下：

$$\theta' = \theta^*, b' = \sigma^b \quad (\text{公式八})$$

则公式七可以转变成

$$P(\theta) = \frac{e^{\theta-\delta}}{1 + e^{\theta-\delta}} \times \frac{\sigma^b}{\sigma^b}$$

$$= \frac{\theta^*}{\theta^* + b^*}$$

$$= \frac{\theta^*}{b^* + \theta^*} \quad (\text{公式九})$$

因此，

$$P(\theta^*) = \theta^* / (b^* + \theta^*) \quad (\text{公式十})$$

公式十即是 Rasch 模式(1960)对成功的机率所下的原始定义。

正确反应机率既如上述定义在 θ^* 量尺上的 $P(\theta^*)$ 所示，不正确反应机率则为

$$Q(\theta^*) = 1 - P(\theta^*), \text{ 亦即是}$$

$$Q(\theta^*) = b^* / (\theta^* + b^*) \quad (\text{公式十一})$$

因此，成功的胜算(odds for success)O 可以定义成

$$O = P(\theta^*) / Q(\theta^*)$$

$$= \theta^* / b^* \quad (\text{公式十二})$$

假设有两位考生在某一试题上的能力各为 θ_1^* 和 θ_2^* ，且其成功的胜算各为 O_1 和 O_2 ，则他们的成功的胜算比为

$$\frac{O_1}{O_2} = \frac{\theta_1^* / b^*}{\theta_2^* / b^*} = \frac{\theta_1^*}{\theta_2^*} \quad (\text{公式十三})$$

公式十三意谓着，在 θ^* 量尺上，若某考生的能力是另一考生能力的两倍，则他答对某一试题的机率也是另一考生的两倍。同理，若同一考生在两题不同难度值(如 b_1^* 和 b_2^*)

的试题上成功的胜算各为 α_1 和 α_2 ，则该考生答对该二试题的胜算比为

$$\frac{\alpha_1}{\alpha_2} = \frac{\theta^* / b_1^*}{\theta^* / b_2^*} = \frac{b_2^*}{b_1^*} \quad (\text{公式十四})$$

由公式十四可以知道，假设第二题试题的难度是第一题难度的两倍（如： $b_2^* = 2b_1^*$ ），则该考生答对第一题较简单的试题的机率是他答对第二题较困难试题的两倍。

上述 θ^* 和 b^* 量尺所具有的比率量尺的特性，仅适用在一个参数模式里。关于二个参数和三个参数模式，则量尺又必须另外定义，有兴趣的读者可自行参考 Hambleton & Swaminatian(1985) 的详细说明。

一个参数模式中另外一种较有意义的非直线转换，便是采「对数胜算」(log-odds)的转换。例如，两位考生对同一试题的成功的胜算比为

$$\frac{\alpha_1}{\alpha_2} = \frac{\theta_1^*}{\theta_2^*} = \frac{e^{a_1}}{e^{a_2}} = e^{(a_1 - a_2)} \quad (\text{公式十五})$$

公式十五取自然对数后，则变成

$$\ln\left(\frac{\alpha_1}{\alpha_2}\right) = (a_1 - a_2) \quad (\text{公式十六})$$

如果两位考生的能力相差一个单位，即

$$\ln(\alpha_1 / \alpha_2) = 1$$

则

$$\alpha_1 / \alpha_2 = e^1 = 2.718$$

亦即，在能力量尺上相差一个单位，则相当于在 θ 量尺上的成功的胜算相差约 2.72 的量。同样的道理，如果同一考生回答两个不同难度的试题，则

$$\ln\left(\frac{\alpha_1}{\alpha_2}\right) = (b_2 - b_1) \quad (\text{公式十七})$$

亦即，在试题难度上相差一个单位，即相当于在成功的胜算上相差约 2.72 的量。

在对数胜算量尺上的单位，即称作「洛基」(logits)。洛基单位可以由下列程序直接求得，亦即

$$\frac{P(\theta)}{Q(\theta)} = e^{b-\theta} \quad (\text{公式十八})$$

则取自然对数后，公式十八的单位即是洛基：

$$\ln\left(\frac{P(\theta)}{Q(\theta)}\right) = \theta - b \quad (\text{公式十九})$$

转换成真实分数量尺

其实， θ 量尺最主要的转换用途是将它转换成真实分数量尺(true-score scale)；因为真实分数量尺的范围是由 0 到 n ， n 为测验的题数，而 θ 量尺的范围却是介于正负无穷大之间（亦即 $-\infty < \theta < \infty$ ），若将 θ 量尺转换成真实分数量尺，不仅有助于我们陈报考生的能力高低，更有助于我们解释测验分数和作为对换测验(test equating)之用。

前面曾经说过，真实分数是答对题数和分数之期望值，以数学公式表示如下：

$$t = E(X) = E\left(\sum_{j=1}^n U_j\right) \quad (\text{公式二十})$$

其中， t 为真实分数， X 为答对题数和分数， U_j 代表第 j 个试题上的反应分数（即答对者给 1 分，答错者给 0 分）， E 代表求期望值的运算符号。若根据期望值的运算方法，公式二十可以展开如下：

$$t = E\left(\sum_{j=1}^n U_j\right) = \sum_{j=1}^n E(U_j)$$

$$= \sum_{j=1}^n [1 \cdot P_j(\theta) + 0 \cdot Q_j(\theta)]$$

$$= \sum_{j=1}^n P_j(\theta) \quad (\text{公式二十一})$$

亦即，真实分数即是能力为 θ 的考生在一堆试题上的试题特征曲线(item characteristic curves)之和。由此看来，真实分数其实就是考生在某一测验上的测验

特征曲线(test characteristic curves)，当然，这种说法也仅有在试题反应模式适用于该数据的条件下才成立。

真实分数可以被看成是 θ 的一种非直线转换，因为 θ 与 τ 间具有一种依序递增的函数关系。另一种常用的转换，便是将 τ 转换成真实比例正确分数(true proportion correct score)或内容范围分数(domain score)如下：

$$\pi = \tau/n = \sum_{j=1}^n P_j(\theta)/n$$

(公式二十二)

可见 π 的值介于0与1之间，如同百分比一般介于0%到100%之间。在一个参数和二个参数模式下， π 的下限值为0；而在三个参数模式时，由于 θ 趋近于 $-C_1$ ，所以 $P_j(\theta)$ 趋近于最低的渐近线 C_j ，故 π 的下限值为 $\sum C_j/n$ ，与之相对的 τ 的下限则为 $\sum C_j$ 。

将 θ 转换成真实分数或内容范围分数有许多好处：第一，负的分数可以被消除，便利于大众的理解能力；第二，新量尺的范围介于0与 n 之间(或0%到100%之间)，分数本身即具有解释涵义在里面；第三，内容范围分数比 θ 量尺更好决定区别精熟与否的切割分数(cut-off score)，便利于精熟测验(mastery testing)的实施；第四，将真实分数对照其相对应的 θ 值，画成一个双向度的分布图，有助于判定切割分数的位置。

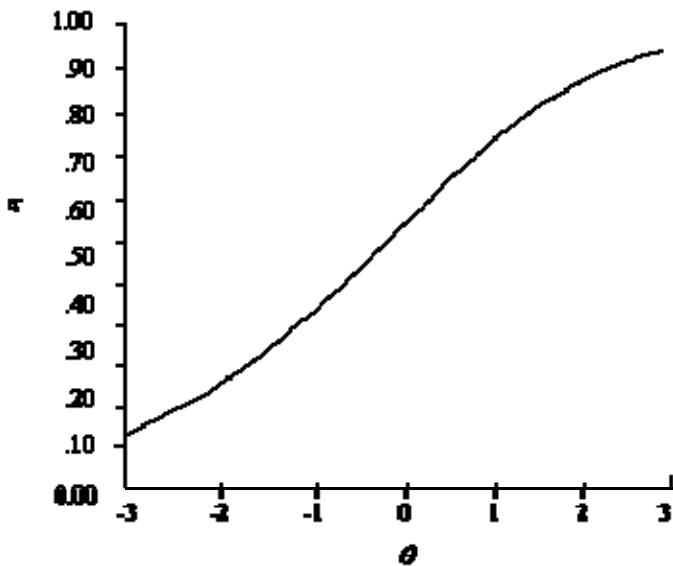
为了说明起见，兹举表一的五个试题的基本数据为例，分别计算 $\theta = -3, -2, -1, 0, 1, 2, 3$ 时，三个参数模式的正确反应机率，并合并这些机率值为真实分数，及算出其相对应的内容范围分数，并画出内容范围分数与 θ 值的分布图，如表二及图一所示。

表一 五个试题的试题参数值

试题	难度	鉴别度	猜测度
1	-2.00	0.80	0.00
2	-1.00	1.00	0.00
3	0.00	1.20	0.10
4	1.00	1.50	0.15
5	2.00	2.00	0.20

表二 θ 、 τ 与 π 之间的关系

θ	$P_1(\theta)$	$P_2(\theta)$	$P_3(\theta)$	$P_4(\theta)$	$P_5(\theta)$	$\tau = \sum P_k(\theta)$	$\pi = \tau' \pi$
-3	.20	.03	.10	.15	.20	.69	.14
-2	.50	.15	.11	.15	.20	1.12	.22
-1	.80	.50	.20	.16	.20	1.85	.37
0	.94	.85	.55	.21	.20	2.75	.55
1	.98	.97	.90	.58	.22	3.65	.73
2	.99	.99	.99	.94	.60	4.51	.90
3	1.00	1.00	1.00	1.00	.96	4.96	.99



图一 θ 与 π 关系之分布图

由图一可以看出 θ 与 π 之间具有依序递增的关系 (monotonically increasing relationship)，而与 π 相对应的 θ 量尺上的分数，即可作为判断精熟与否的切割分数 *Hambleton & deGruiter, 1983)。

参考书目

1. Hambleton, R. K., & deGruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement*, 20, 355-367.
2. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: kluwer.

3. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.
4. Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.
5. Woodcock, R. W. (1978). Development and standardization of the Woodcock-Johnson Psycho-Educational Battery. Hingham, MA: Teaching Resources Corporation.
6. Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

本文转载自研习信息 9 卷 (5 期), 8-12 页

第七章

試題反應理論的介紹(七) 訊息函數 (Information Functions)

政大教育系教授 余民宁 着

我们曾在前几篇文章里谈到过，试题反应理论与古典测验理论有两点不同：一为参数具有不变性(invariance)，另一为讯息函数(information function)概念的提出。不变性已经在前几篇文章里谈论过了，本文即集中在讯息函数的讨论上。

基本概念

试题反应理论提出一个能够用来描述试题或测验、挑选测验试题、以及比较测验的相对效能的实用方法，该方法即需要使用试题讯息函数(item information function)，作为建立、分析、与诊断测验的主要参考依据。试题讯息函数的定义如下：

$$I_i(\theta) = \frac{[P_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad i = 1, \dots, n \quad (\text{公式一})$$

其中的符号， $I_i(\theta)$ 代表试题*i* 在能力为 θ 上所提供的讯息， $P_i(\theta)$ 为在 θ 点上的

$P_i(\theta)$ 值的导数，而 $A_i(\theta)$ 为能力 θ 在试题*i* 上的试题反应函数， $Q_i(\theta) = 1 - P_i(\theta)$ 。

试题讯息函数可以应用到前面所谈到的一个、二个、与三个参数对数形试题反应模式，这些模式都适合用于二分法计分(dichotomously scored)的测验资料。例如，以三个

参数对数形模式为例，公式一可以化简为(Birnbaum, 1968; Lord, 1980):

$$I_i(\theta) = \frac{a_i^2(1-C_j)}{[C_j + e^{a_i(\theta-b_i)}][1 + e^{-a_i(\theta-b_i)}]^2} \quad (\text{公式二})$$

从公式二里，我们很容易便可推知 a , b , 和 c 参数在试题讯息函数中所扮演的角色：(1)当 b 值愈接近 θ 时，讯息量较大；反之， b 值愈远离 θ 时，讯息量则较小；(2)当 a 参数较高时，讯息量也会较大；(3)当 c 参数接近 0 时，讯息量则会增加。

试题讯息函数在测验的发展与编制上，以及试题好坏的诊断上，扮演着举足轻重的角色，因为它能表示出试题对能力估计正确性的贡献量大小。该贡献量的大小，端受两个主要因素的决定：一为试题的鉴别度参数的大小（亦即， a 值愈大，试题特征曲线便愈陡， $R(\theta)$ 的斜率便愈大，所以讯息量便愈高）；另一为试题的难度参数，它的位置会决定讯息量的高低。Birnbaum(1968)指出，某个试题所提供的最大讯息量，刚好出现在能力参数为 θ_{\max} 的点上， θ_{\max} 的值为：

$$\theta_{\max} = b_i + \frac{1}{a_i} \ln [0.5(1 + \sqrt{1 + 8C_j})] \quad (\text{公式三})$$

如果猜测机率为最小时（亦即，当 $C_j = 0$ 时），则 $\theta_{\max} = b_i$ 。一般而言，当 $C_j > 0$ 时，某个试题在能力水平比其难度值稍高的位置上，所提供的讯息量会达到最大。讯息量最大值所对应的能力水平，即代表该试题所能最精确测量或估计到的能力参数估计值。因此，算出试题的最大讯息量，便可知道该试题所精确测量到的潜在特质大概是多少，或者是说该试题适合何种潜在特质程度的测量。

在发展测验或评鉴试题上若使用试题讯息函数的协助，尚需有个基本前提必须先成立，那就是假设我们所选用的试题特征曲线(ICC)能够适用于测验数据。如果这种数据与试题特征曲线间的适合度很差的话，则我们所计算得到的试题参数估计值和试题讯息函数，将会产生误导的作用；甚至，当这个适合度尚属良好时，如果 a 参数很低，且 c 参数很高，则试题的有用性亦会受到限制，它无法通用于所有的测验中。此外，测验试题的有用性有时也受到测验编制者在编制某种具有特殊用途测验的需求的限制。因此，某个试题在某种能力量尺上也许可以提供相当可观的讯息量，但在另一种用途的能力量尺上，则无法提供丝毫无价值的讯息量。

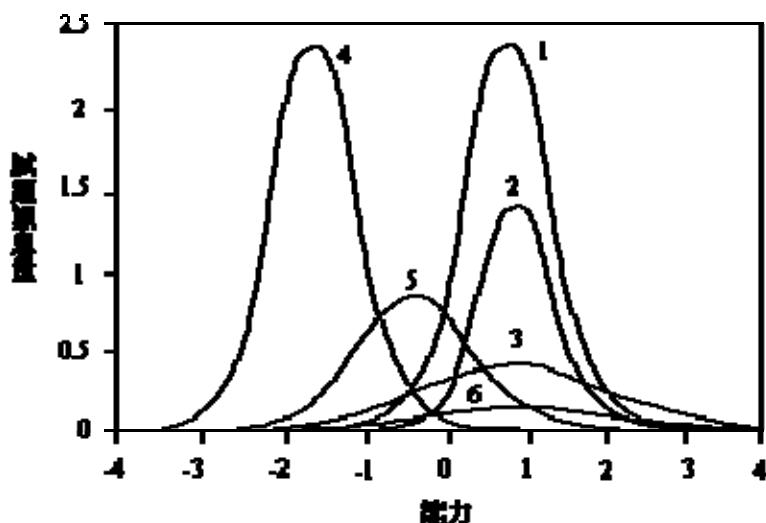
有了上述基本概念后，我们举六个不同讯息量的试题为例，说明讯息量所具有的应用涵义。

讯息函数的例图

试题讯息函数与能力水平二者是组成讯息函数图的两个主轴，所画出的讯息函数就如图一所示，它是根据表一的试题参数值所画成的。

表一 六个试题的试题参数值

测验试题	试题参数		
	b_i	a_i	c_i
1	1.00	1.80	0.00
2	1.00	0.80	0.00
3	1.00	1.80	0.25
4	-1.50	1.80	0.00
5	-0.50	1.20	0.10
6	0.50	0.40	0.15



图一 表一中六个试题的试题讯息函数

由图一所示，我们可知它们提供许多宝贵的见解：

1. 当 $C > 0$ 时，试题所提供的最大讯息量，大约出现在它的难度水平或比其难度水平稍大的位置（我们只要比较最大讯息量所对应的能力量尺上的位置和表一中的相对应的 b 值便知）。
2. 试题的鉴别度参数很显然地影响试题所提供的讯息量（这点可由比较试题一和试题二的试题讯息函数中得知）。
3. 在其他条件均相等的情况下，具有 $C > 0$ 的试题比较不适于用来评定能力水平（这点可由比较试题一和试题三的试题讯息函数中得知）。
4. 具有较低鉴别力的试题，在整份测验中则几乎不具有任何统计学的用处（如试题六一般）。
5. 在评定某些能力水平的范围内，即使是最具有鉴别力的试题（如试题一和试题

四），也会比某些鉴别力较差的试题（如试题五），提供较少的讯息量。例如：在评定具有中等能力的考生能力（即能力水平约在-.50 左右者）时，试题五比试题一和试题四提供较为有用的讯息；换句话说，对中等能力的考生而言，试题五比试题一和试题四较为适当且有用。

由上述的五个见解可知，试题讯息函数可以提供我们判断测验试题和编制测验的有效性的一个新方向。

一般而言， $C > 0$ 的试题讯息函数都会比 $C = 0$ 的试题讯息函数还小，在这种情况下，研究者也许会考虑使用一个或二个参数模式，以求合适所使用的测验数据。结果所得到的试题讯息函数将会比较高些；因此，也唯有在试题特征曲线能够适用于所分析的数据时，一个和二个参数的试题讯息曲线才能发挥用处。若试题特征曲线并无法很适当地适用于所分析的测验资料，且其相对应的试题讯息曲线也偏离理想的形状很远，而我们仍然使用它们时，则我们会获得具有误导作用的结果。de Gruijter(1986)便曾举例说明，在某些情况下，样本太少时而仍然使用 Rasch 模式，便会产生偏差的结果。

测验讯息函数

根据 Birnbaum(1968) 的推导，一份测验的讯息函数 (test information function) 是指它在某一个 θ 值上所提供的讯息量，该讯息量刚好是在 θ 值上的试题讯息函数之总和，记作 $I(\theta)$ ：

$$I(\theta) = \sum_i I_i(\theta) \quad (\text{公式四})$$

由于在 θ 值上的测验讯息函数是其试题讯息函数之总和，从公式四里可以看出：每个试题都单独地对测验讯息函数作贡献，因此，每个试题所作的贡献量大小，并不受在该测验中其他试题的影响。这个特性是古典测验理论所没有的，也正是试题反应理论所具有两项特点之一。然而，测验试题对测验信度和试题鉴别度指标（如：点二系列相关系数）的贡献，却受在该测验中其他试题特性的影响，而无法单独地决定；因为在计算这些指标时必须用到测验分数，而测验分数却依所选择的测验试题的不同而不同。甚至，只要改变一个试题，便会对测验分数产生影响，紧接着，古典的试题和测验指标也会随着改变。在 θ 值上的测验讯息量与该能力的估计值的精确性成平方根反比，其符号记作：

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (\text{公式五})$$

其中， $SE(\hat{\theta})$ 称作估计标准误 (standard error of estimation)。该项指标只要在能力参数的最大近似估计值求出后，便可计算得出。有了能力参数的最大近似估计值，并且也求出在 θ 值上的测验讯息之后，我们便可以估计信赖区间的方式来解释能

力估计值的涵义。一般而言，最大的测验讯息量所对应的能力估计值 θ ，便是该份测验所精确测量到的能力参数，也可以说是该份测验适用于该能力估计值范围内的测量。

有关这点说明，我们可以由公式五中的定义得知，当 $I(\theta)$ 值达到最大时， $SE(\hat{\theta})$ 值便达到最小，也就是说该 θ 值的最大近似估计值的估计误差达到最小，亦即此时的 θ 的最大近似估计值最精确。

在试题反应理论的架构里， $SE(\hat{\theta})$ 所扮演的角色和古典测验理论中的测量标准误(standard error of measurement)的角色相同，然而有一点需要注意者， $SE(\hat{\theta})$ 的值随着能力水平的不同而不同，但古典的测量标准误对所有能力水平的考生而言，却都是一致的；换句话说，古典的测量标准误的意义是认为每位考生能力估计值的误差都是一致的，而试题反应理论的估计标准误则认为每位具有不同能力水平的考生，皆应有不相同的估计误差（或估计的精确性）。

其实， θ 的最大近似估计值 $\hat{\theta}$ 的标准误， $SE(\hat{\theta})$ ，是这个特定 θ 值的最大近似估计值所构成的渐近性常态分配的标准偏差。当测验的长度够长时，该分配是呈常态的；即使是测验长度仅有 10 至 20 个试题，这种以常态分配的估计方法，也可以满足多数测验目的的要求(Samejima, 1977)。

一般而言，估计标准误的大小受三个因素的影响：(1)测验试题的数目（例如：较长的测验会有较小的标准误）；(2)测验试题的质量（例如：鉴别度较高的试题往往让能力低的考生没有侥幸猜对的机会，所以它的标准误便较小）；(3)试题难度与考生能力之间的配合程度（例如：组成测验的试题难度参数若与考生的能力参数相近，则其标准误会比具有相当困难或相当简单试题的测验的标准误还小）。标准误的大小很快地会趋近于稳定，因此，当讯息量增加到超过 25 时，讯息函数对能力估计值的估计误差的影响，仅会发生小小的作用，典型的例子可以参见 Green, Yen, & Burkett(1989)的论文。

相对的效能

有了测验讯息函数之后，测验编制学家们往往感兴趣的是：比较两份或多份测量到同样能力的测验讯息函数。比较两份或多份测验的讯息函数，可以提供测验专家作测验评鉴和选择的参考（参见 Lord (1977)的例子）。所以在发展一份全国性的成就测验时，往往就需要比较不同测验的讯息函数，以帮助选择优良试题来组成所需的测验；或者，在编制一份标准化成就测验时，可参考过去有关学生表现的讯息函数概况，再优先挑选在某段能力范围内能产生最大讯息量的试题，汇编成我们所需的标准成就测验（至于其他因素，如：效度、成本、内容、和测验长度等，当然也必须在考虑之内）。

比较两份测验的讯息函数是这样进行的：把两份同样测得能力估计值为 θ 的测验讯息函数相除，该商值便定义为某个测验的相对效能(relative efficiency)：

$$RE(\theta) = \frac{I_A(\theta)}{I_S(\theta)} \quad (\text{公式六})$$

其中， $RE(\theta)$ 便是相对效能，而 $I_A(\theta)$ 和 $I_B(\theta)$ 则为定义在一个共同能力量尺 θ 上的 **A** 测验和 **B** 测验的讯息函数。相对效能的涵义可由下列例子的说明得知：假设 $I_A(\theta) = 25.0$, $I_B(\theta) = 20.0$ ，则代入公式六，得 $RE(\theta) = 1.25$ ，我们可以解释为：

「在能力水平为 θ 时，测验 **A** 所发挥的效能比测验 **B** 所发挥的效能要多（或长）25%，因此，测验 **B** 必须要加长 25%（即把讯息函数相当的试题加入原有的测验试题中），才能产生与测验 **A** 对 θ 值一样的精确测量；或者是，测验 **A** 可以缩短 20% 的长度，就可以产生与测验 **B** 对 θ 值一样精确的能力估计值。」当然，上述解释中的加长或缩短测验长度的作法，都是假设所增减的试题都和原有测验中的试题，一样具有可资比较的统计质量（如：类似的难度、鉴别度，产生大约一致的讯息量，都适用于同一范围程度内的 θ 值的测量等）。

由上述的举例说明，讯息函数的应用性非常的广，我们将在后续文章里逐一介绍试题和测验函数，以及相对效能的应用实例，尤其是应用在测验编制里。

参考书目

1. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (chapters 17-20). Reading, MA: Addison-Wesley.
2. De Gruijter, D. N. M. (1986). Small N does not always justify the Rasch model. Applied Psychological Measurement, 10, 187-194.
3. Green, D. R., Yen, W. M., & Burkett, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2(4), 297-312.
4. Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.
5. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
6. Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

本文转载自研习信息 9 卷 (6 期)，5-9 页

第八章

試題反應理論的介紹(八)

....測驗編製
(Test Construction)

政大教育系教授 余民宁 着

在古典测验理论下，编制成就或性向测验的方法，往往仅考虑试题的内容和特征（如：难度和鉴别度），就当成是选择试题的依据；例如：先挑选出鉴别度较高（如：大于.25）的试题，再依据实施测验的目的和考生的能力分配情况，挑选出难度较适中的试题，编成整份测验（郭生玉，民 79，页 269-272）。

然而，我们也在前几篇文章里评论过古典测验理论所使用指标的缺失，例如：难度和鉴别度都不是不变值(*invariant*)，它们会随着考生群体的能力分配的不同，而有不同的估计值出现，这些估计值都是样本依赖(*sample dependent*)的估计值。因此，用来决定试题指针的样本组能否适切代表测验所要测量的母群体，便成为决定某种审慎选择试题的技术能否成功的主要因素。当这个代表性堪疑时，所获得的试题指标（如：难度和鉴别度）便不太适用于将来所欲测量的母群体。

此外，由于学生生理与心理的成熟与成长，原本在学年度开始所建立的试题指标，到了学年度终了时，便不适用于原本的学生族群，因为参与测验的学生的能力分配，经过一学年的期间，已发生明显的变化，因此导致期初所建立的测验试题，无法适当地应用到期末的测验情境中。

另一种情况也会使得古典的试题指标无法适用于未来所欲测量的母群体，那就是来自题库(*item bank*)的测验编制。在发展一套题库之时，所有要被放入题库的试题特征，应该都已经事先被估算出来，并且事先决定好。实际上，这些被称作「实验性」的试题，是在被编入一份测验卷，并对一群受试者施测后，才计算出试题指标估计值的。由于实验性试题的数目远比测验卷数还多，我们只能把它们编成几份测验卷，每份均含有不同的实验性试题和不同的题型，再拿来对不同族群的受试者施测。由于我们无法保证这些接受不同题型测验的学生，都是能力相等的学生，因此，我们在不同族群受试者下所建立起来的试题指标，彼此间便无法比较。在这种情况下，题库试题的指标若被假设成是可以比较的，则从该题库中所建立起来的任何测验，便无法适合用于某一特定的群体。

除了试题指标本身不具有不变性之外，即使在已有一份编制良好的现成题库下，古典测验理论的测验编制方法，仍有一项很严重的缺失，那就是被选入编成测验的试题，无法满足事前订定的测量精确度的要求。试题对测验信度的贡献量，不仅受该试题特征的影响，同时也受到该试题与其他试题间关联性的影响。因此，我们无法单独计算某个试题对测验信度，甚至对测验的测量标准误的贡献量，而不受其他试题的影响。

为了弥补古典测验理论在编制测验上所面临的困难和缺失，试题反应理论提出一项比较强而有力的方法来克服这种窘境，那就是运用试题和测验讯息函数来参与编制测验的工作。运用试题与测验讯息函数的最大好处是，它可以挑选出对满足某份特殊测验所需的讯息总量最有贡献的试题，以编制成可以达成测量目标的测验卷。因为，讯息量和测验的精确度息息相关，并且，试题难度指标和学生能力指标又定义在同一量尺上，所以，我们可以在任何能力水平上，挑选出最能精确测量（亦即该测量标准误差最小）到该能力范围的试题，以编制成我们所需要的测验。

测验编制的基本方法

试题反应理论应用到测验编制上，最常用的工具莫过于使用讯息函数(*information function*)。根据一般建立题库的过程，在选定合适的试题反应模式来分析数据后，除了可以获得试题参数和学生的能力参数估计值外，也可以获得讯息函数值。利用试题讯息函数，以编制能够满足某种特殊需求的测验编制过程，已由学者

(Lord, 1977)提出纲要如下：

1. 决定所要的测验讯息函数的形状，该形状的曲线便叫作「目标讯息函数」(target information function)。
2. 由题库中先挑选一组试题，使得这些试题的试题讯息量累加起来的和，能够填满目标讯息函数下最难填的部份（通常是讯息函数曲线最突起的部份）。
3. 每加入一个试题，便计算现有测验试题所有的测验讯息函数。
4. 继续上述的选题步骤，直到测验讯息函数接近目标讯息函数到达某种令人满意的程度为止。

上述这些测验编制的步骤，通常需要仰赖大计算机和测验编制专家的共同合作，否则光靠笔算会费时、费力。

由已知（或现成）的试题反应模式下所建立起来的题库中，我们可以根据 Lord(1977)所提出的纲要，编制出可以在某个能力范围内充分发挥鉴别功能的测验来；也就是说，假设我们已知某组受试者们的能力水平，我们便可以挑选出能够使该能力范围内的测验讯息量达到最大的测验试题来，以作为测量该等能力水平的工具。这种挑选测验试题的作法，将能增进对能力参数估计值的精确性。

举例来说，根据 Lord(1977)所提的纲要，一个涵盖范围较广的能力测验(broad-range ability test)，其目标讯息函数应该是个相当平坦的曲线，它所表示的涵意是，在整个能力量尺上，该测验希望能够提供几乎是同样精确的能力估计值，以表明它所能适用的能力范围较为宽广。而对一个设有切割分数(cut-off score)以区别精熟者(masters)和非精熟者(nonmasters)的效标参照测验(criterion-referenced test)而言，其所期望获得的目标讯息函数，应该是个对应于能力量尺上的切割分数附近，呈现极为尖狭峰分配的曲线，这种情况显示出，在切割分数附近，该测验最能够精确测量到区分精熟与非精熟二者的能力估计值。

透过试题讯息函数的使用，测验编制者可以编制出满足各种特殊需求的测验来。例如，Yen(1983)便曾举例说明，如何运用试题讯息函数来编制一份大规模的测验。van der Linden & Boekkooi-Timmeringa(1989)也已发展出一套程序，说明在测验上加诸一些限制，以确保内容效度、适当的测验长度、和其他特征之后，可以自动挑选测验试题以符合某种测验讯息函数的作法。

为了说明上述的过程起见，兹举一份成就测验为例。就成就测验而言，前测时的表现情形往往远低于后测的表现情形，这是一种很常见的现象。有鉴于此，测验编制者便可以挑选较为简单的试题作为前测的内容，而挑选较为困难的试题作为后测的内涵。在每一个测验情境里，考生能力范围所最常出现的地方，其测量的精确性往往将达到最大。甚至于，由于在这两份测验上的试题，都是在测量相同的能力，而且，能力估计值也不受特别挑选的试题的影响，因此，后测的能力估计值减去前测的能力估计值，其差值便可以用来测量成长(growth)量的大小。

de Gruijter & Hambleton(1983)和 Hambleton & de Gruijter(1983)已着手研究，在测验编制之前便先决定好切割分数或测验的通过标准，看看最理想的试题挑选方法，会对一份测验的决策正确性产生什么样的影响。为了解释这项结果，通常是以随机的方式来挑选试题，以编制成所需要的测验。在效标参照测验的编制过程中，从一堆现成的候选测验试题库里，以随机方式挑选试题以组成测验，是一种常用的作品。只不过是，依随机方式所挑选出的测验试题所组成的测验，其错误率(error rates)（亦即是造成分类错误的可能机率）几乎是依最理想方式来挑选测验试题以组成测验

所造成的错误率的两倍。因此，以试题反应理论为架构，来挑选最理想的测验试题的作法是有可能的，因为试题、学生、和切割分数都是建立在同一量尺的基础上，所以方便测验的编制与测验结果的解释。

其实，设定目标讯息函数和挑选试题的程序，仍存在有许多值得商榷的问题。其中一个便是，单依靠统计学的效标来挑选试题的作法，并没有办法保证就可以编制出一个内容有效(content-valid)的测验来。可惜，我们通常却过度强调统计学的效标，而忽略试题内容在测验编制上所扮演的重要角色。忽视内容的考虑事项，往往会导致编制出一个缺乏内容效度的测验来。为了解决这个难题，van der Linden & Boekkooi-Timmeringa(1989)使用线性规划(linear programming)的技术，提出许多同时考虑试题内容和统计学的效标等有用的方法，以作为挑选试题的参考依据。

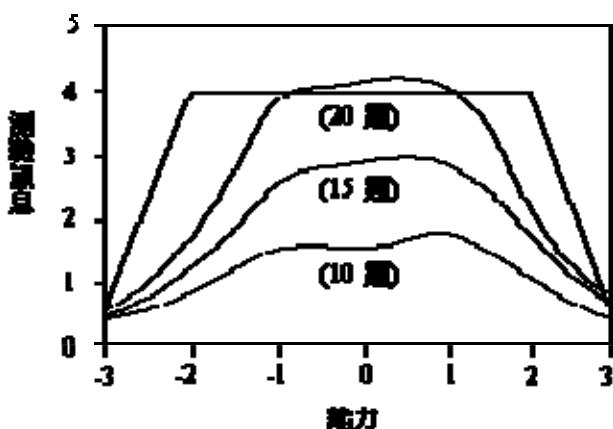
使用试题讯息函数来作为测验编制的依据，还有另一项问题会产生，那就是很可能高估高鉴别度（即 α ）值，以致于讯息函数也许会产生偏差。使用具有高鉴别度的试题所编制出的测验，很可能会与期望中的测验相去甚远。由于测验讯息函数将会被高估，所以增加额外的几个试题到测验里，也许会缓和高估的情形。而最好的解决办法，还是尽量使用大样本，以确保试题参数的估计值都很正确、很稳定。

下列所举的例子，是说明如何运用讯息函数来编制特殊测量目的的测验。

广泛能力测验的编制

假设某位测验专家想编制一份包含广泛能力的测验，他认为该能力范围应涵盖 **(-2.00, 2.00)** 之间，并且只容许有 .50 以下的估计标准误存在，而在此能力范围外者

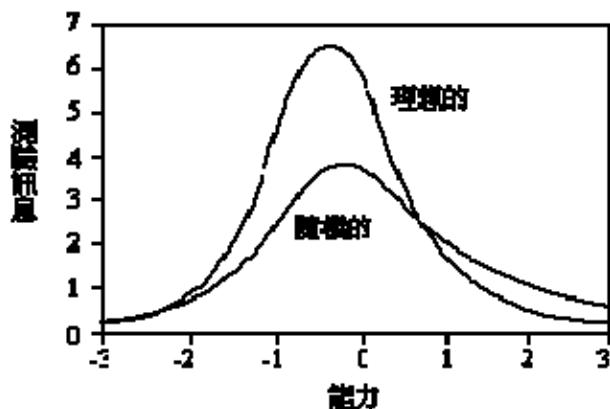
则允许有少许较大的误差存在。例如，假设选定 $SE(\theta) = .50$ ，则 $I(\theta) = 4.0$ ，典型的目标讯息函数可如图一所示建立起来。为了编制一份能够满足此目标，并且具有愈少试题愈好的测验，我们就必须从具有难度值介于 -2.00 和 2.00 之间、高鉴别度、和低猜测度的试题群中，去挑选符合要求的候选试题。图一所示，即为在既定的目标讯息函数（即 θ 值介于 **± 2.0** 之间，且 $I(\theta) = 4.0$ ，呈现平坦的曲线）下，从题库中挑选出最理想的 10、15、和 20 题测验试题后，所计算出的测验讯息函数。很明显的可以从图一看出，20 题下的测验，最为接近我们想要编制的目标测验。若增加难度值接近 **± 2.0** 的试题，则所获得的测验讯息函数更加接近目标讯息函数。



图一 含有 10、15、和 20 题试题测验的测验讯息函数

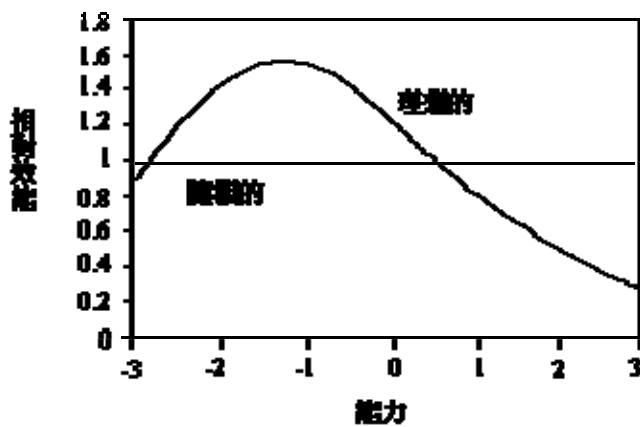
效标参照测验的编制

假设某位测验专家想要编制一份含有 15 个试题的效标参照测验，使得其测验讯息函数在切割分数 $\theta = -.50$ 处达到最大。为了比较起见，也以一般常用的方法随机抽取 15 个试题编成测验（称作标准测验），并计算出其应有的测验讯息函数。兹将这两种不同方式挑选试题编制成的测验讯息函数，画于图二里，以资比较。



图二 理想的和随机的挑选方法下 15 个试题的测验讯息函数

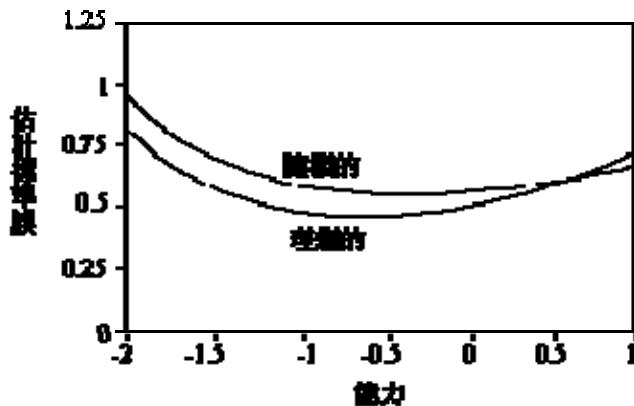
图三所示，为理想的测验对标准测验之相对效能图。很明显的，理想的测验在切割分数（即 $\theta = -.50$ ）处，提供较大的测量精确性；它比标准测验在此处高出 60% 的相对效能，也就是说，标准测验的长度必须从 15 题增加到 24 题，才能发挥与理想的测验同等的效能。



图三 理想的对随机的 15 题试题测验的相对效能

由图二和图三可以看出，对高能力学生而言，理想的测验表现的不如标准测验表现的好。这是由于理想的测验仅包含能够在切割分数附近发挥鉴别功能的试题，而忽略许多适合于高能力学生的试题的缘故。由此可见，标准测验包含比较多的异质试题在内。

实际说来，题库中的试题愈异质化，或所欲编制之测验长度占题库大小的比率愈小，则理想的试题挑选方法远比随机的试题挑选方法较优。相对于这两种挑选方法下的测验讯息函数的估计标准误，则如图四所示。由图四可知，理想的测验的估计标准误比随机的测验的估计标准误还小。



图四 理想的与随机的挑选方法下 15 个试题的估计标准误

参考书目

1. 郭生玉（民 79）。心理与教育测验（五版）。台北：精华。
2. de Gruijter, D. N. M., & Hambleton, R. K. (1983). Using item response models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), Applications of item response theory (pp.142-154). Vancouver, BC: Educational Research Institute of British Columbia.
3. Hambleton, R. K., & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20, 355-367.
4. Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.
5. Van der Linder, W. J., & Boekkooi-Timmeringa, E. (1989). A maximum model for test design with practical constraints. Psychometrika, 54, 237-247.
6. Yen, M. W. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 123-141). Vancouver, BC: Educational Research Institute of British Columbia.

第九章

試題反應理論的介紹(九) 測驗分數的等化(上) ... (Test score equating)

政大教育系教授 余民宁 着

在许多测量到相同能力的不同测验中，其间测验分数的比较性如何？一直是测验编制者、测量专家、以及接受测验的人，所一致关心和重视的问题。例如：两位考生接受两种不同的测验，其测验分数该如何比较？这个问题当遇到如证照考试、甄试入选、或及格与不及格的决定时，会更形重要，因为这些决定不应该受考生接受不同测验的影响，而是各种测验分数间该如何客观、有效地进行比较的问题。

要比较在不同测验（假设为 X 和 Y ）上所获得的测验分数，我们必须先建立起这两种测验间的等化分数(equating scores)的程序。透过这个程序，我们可以建立起 X 和 Y 分数间的一种对等关系，而可以把 X 测验上的得分转换成 Y 测验上的量尺分数。

因此，某位考生在 X 测验上得 x 分，可以换成 Y 测验的 y 分，这个分数就可以拿来和某生在 Y 测验上所得的 y 分作比较。当面临作证照考试、甄试入选、或及格与不及格的决定时，在 X 测验上的切割分数(cut-off score) x_c 也可以转换成 Y 测验上的切割分数 y_c ，而这种经过转换过的切割分数，就可以帮助教师或研究者针对某生在 Y 测验上的得分，作出适当的决定。本文即在讨论不同测验分数间如何等化的问题。

试题反应理论的衔接方法

古典测验理论所使用的等化方法，大致可以归成两类：相等百分比等化法(equipercentile equating)和直线等化法(linear equating) (Angoff., 1971; kolen, 1988)。然而，这些等化方法通常无法满足公平(equity)的需求。Lord(1977, 1980)认为测验分数的等化，不应该受某一特定能力程度的考生接受 X 测验或 Y 测验的不同影响。测验分数要能公平的等化，必须满足五项需求：

1. 测量不同特质的测验无法等化；
2. 在信度不相等的测验上的原始分数无法等化；
3. 在难度变化大的测验上的原始分数无法等化，因为在不同的能力水平上的测验不会具有相等的信度；
4. 除非 X 和 Y 测验是全然复本测验(strictly parallel tests)，否则，在该二测验上容易错估的分数无法等化；
5. 具有完全信度的测验分数才可以等化。

此外，对称性(symmetry)和不变性(invariance)等二条件，亦是进行测验分数之等化所必备的。对称性是指等化不应该受使用何种测验为参照测验的影响；不变性则指等化的程序应该是样本独立的（亦即指不受所选用样本的影响）。这两者亦是古

典的等化方法所不容易满足的条件。有鉴于此，诉诸理论与方法皆称严谨的试题反应理论，应是解决之道。

我们曾于前文说过，当试题反应模式适合用于某份数据时，由于它具有不变性的特性，所以直接比较两位接受不同测验的考生的能力参数值是可行的。又根据试题反应理论的说法，能力估计值 θ 不受试题集合的不同影响，因此，只要试题参数为已知，两位接受不同部份试题（或不同测验）的考生的能力估计值，便已经建立在一个共同的量尺上。在共同量尺上的参数估计值（不论是 a 或 b 或 θ 值）可径行直接比较，没有加以等化或量化(scaling)的必要。唯一需要进行等化或量化的情况，是当试题和能力参数均为未知时，由于在这种情况下，我们所每次校准的能力和试题参数值的原点（如：平均数）和单位长（如：标准偏差），都是随机决定的，所以根据不同族群考生的反应资料所估计出来的不同试题参数和能力参数值不能直接比较，必须经过衔接(linking)的手续，将参数值转化到同一量尺上时，才可以进行比较。基本上而言，衔接算是等化的一种，只不过它所遭遇的问题较为单纯，受到的学术争议也较少。

假设 A 和 B 代表两组不同的考生。我们可以估计出这两组考生在两份测验上的试题和能力参数值，则除了受随机抽样所造成的误差影响外，这两组参数估计值一定可以满足下列的关系：

$$\theta_A = a\theta_B + b \quad (\text{公式一})$$

$$b_A = ab_B + gb \quad (\text{公式二})$$

$$a_A = a_B / a \quad (\text{公式三})$$

在试题反应理论里，衔接的主要工作，便是在透过线性转换(linear transformation)的关系，有效地决定常数 a 和 b 的作法。在应用上而言，比较有效的衔接方法有下列五种：

1. 同时校准法(concurrent calibration method): 此法假设两份测验中，有部份试题是相同的，或有些考生同时接受这两份测验；亦即两组的反应资料有重迭的部份。在实际作法上，此法乃将这两组的反应数据加以合并后，输入计算机，一起估计试题和能力参数值。由于合并的数据将会很庞大，挑选记忆容量大、指令周期快的计算机，将是成败的决定因素。
2. b 值固定法(fixed b's method): 此法假设两份测验中，有部份试题是相同的。在实际作法上，此法乃先估计第一份测验的试题参数，在估计第二份测验的试题参数时，相同试题所估计出的 b 值不再重新估计，而固定在原来的估计值上。如此一来，第二份测验的试题参数在估计时，就能参考固定的 b 值的量尺，而做适当的调整，以达到衔接的目的。
3. b 值等化法(equated b's method): 此法亦假设在两份测验中，有部份试题是相同的。在实际作法上，此法乃分别估计两份测验的试题参数。因为相同试题的 b 估计值有两组，其平均数为 μ ，标准偏差为 σ ，均可用来估计等化时所

需要的常数项， α 和 β 。其估计式如下：

$$\hat{\alpha} = \sigma_A / \sigma_B \quad (\text{公式四})$$

$$\hat{\beta} = \mu_A - (\sigma_A / \sigma_B) \mu_B \quad (\text{公式五})$$

再将公式四和公式五中的 $\hat{\alpha}$ 和 $\hat{\beta}$ 代入公式一、公式二、和公式三，便可以将第二组的估计值转化到第一组的估计值的量尺上，而达到衔接的目的。

4. 特征曲线法(characteristic curve method): 此法亦假设两份测验中，有部份试题是相同的。并且认为如果估计误差很小，参数值经过衔接手续后，每位

考生在两份测验中相同试题上所得的真实分数(true score)（亦即 $\sum_{i=1}^K R_i(\theta)$ ），

其中 K 为相同试题的数目），必须相等。在实际作法上，此法在估计 α 和 β 值时，先随机选取 N 个 θ^* 值，这些 θ^* 值必须和两组的 θ 估计值无任何关联，接着解出下列函数值的极小化，以便求出 α 和 β 值：

$$F = \sum_{j=1}^N \sum_{i=1}^K [\hat{P}_{Ai}(\theta_j) - \hat{P}_{Bi}^*(\theta_j)]^2 \quad (\text{公式六})$$

其中，

$$\hat{P}_{Bi}^*(\theta) = C_{Bi} + \frac{1 - \hat{C}_{Bi}}{1 + \exp(-\hat{\alpha}_B / \alpha[\theta_j - (\alpha C_{Bi} + \beta)])} \quad (\text{公式七})$$

再将估计出来的 α 和 β 值，利用公式一、公式二、和公式三，将第二组估计值转化到第一组估计值的量尺上，以达到衔接的目的。

1. 最小卡方法(minimum chi-square method): 此法与特征曲线法极为类似，并且把参数的估计误差考虑在内；亦即在解 F 函数极小化时，估计误差大的试题参数，其加权值应较小，而误差小的加权值应较大。根据此看法，

Divgi(1985)建议解下列函数值 Q 的极小化，以求出 α 和 β 值：

$$Q = \sum_{i,j} [W_i^2 (\alpha_{ij} - \hat{\alpha}_{ij})^2 + 2W_{ij}(\alpha_{ij} - \hat{\alpha}_{ij})(\beta_{ij} - \hat{\beta}_{ij}) + W_j^2 (\beta_{ij} - \hat{\beta}_{ij})^2]$$

(公八)

其中的 W 代表加权值，其大小和估计误差有关；而 $\hat{\alpha}_i$ 和 $\hat{\beta}_j$ 是根据公式二和公式三作线性转换后的参数值。如果参数值是以最大近似值估计法(maximum likelihood estimation)估计出来的，则误差值可以用讯息函数的倒数来代替。另一种衔接情况是，当有部份考生重复接受两份测验，则衔接的工作也可以根据考生在两份测验上 θ 值的平均数和标准偏差来估算 α 和 β 值，唯独这种根据 $\hat{\theta}$ 值来估计 α 和 β 值的衔接效果不佳，即使先将误差过大的 $\hat{\theta}$ 删去，再估计 α 和 β 值，其改进效果亦仍有限(Liou, 1990)。

衔接法的设计

在许多情况下，衔接的工作只是在将两份或多份测验的试题参数估计值，放置在一个共同的量尺上而已。这项举动不仅使不同测验之难度水平得以比较，更能够促进试题题库(item bank)的发展(Vale, 1986)。不过，欲将两份测验的反应数据转换到同一量尺上时，我们必须藉助相同试题或重复考生。如果测验在发展之初，即已考虑到为将来的衔接作准备，则在施测或试题的设计上，就必须注意此衔接的原则。简单地分，下列四种衔接法的设计，可以使试题参数（或其估计值）得以转换到共同的量尺上：

1. 单一组设计(Single-group design)：将欲衔接的两份测验，给予同一组考生施测。这种作法最简单，但最不实际，因为施测时间会延长，考生个人的身体疲劳或重复练习的因素，都会影响到参数的估计和衔接的结果。
2. 相等组设计(Equivalent-group design)：将欲衔接的两份测验，给予随机选择出来的相似但不完全相同的两组考生施测。此法较为实际，并且可以避免受疲劳和练习等因素的影响。
3. 定锚测验设计(Anchor-test design)：将欲衔接的测验给予两组不同考生施测，每组考生另接受一部份共同试题的测验（可能是附属于每一份测验里，或额外附加的试题皆可），称作定锚测验(anchor test)。此法最为常用，并且，如果定锚试题选得好的话（参见 Klein & Jarjoura, 1985），此法可以避免单一组或相等组设计所遭遇到的问题。
4. 共同考生设计(Common-person design)：将欲衔接的两份测验给予两组不同考生施测，其中有一部份考生重复接受这两份测验。由于共同考生所接受的测验是加倍的份量，所以这种设计也会遭遇同单一组设计一样的缺失。

其实，上述的设计方法是较简略的归类，若予以细分，收集数据以便进行衔接的设计方法，可参见图一所示。兹扼要分述第 2 种以后的设计如下：

(2) 定锚测验随机分组设计：此法与前述定锚测验设计相同，唯此定锚测验的试题内容

必

须与原二份测验十分类似，且测验长度相当于一个分测验。通常定锚测验都暗藏在大

测验中，使考生不易察觉。三份测验同时进行校准，或根据定锚测验资料来估计 α 和 β 值。

(3) 定锚测验不等组设计：前法是以随机方式来混合这两份测验，所以考生接受那一份测

验的机会是一样的。此法乃实行：一组考生在学期中考第一份测验，另一组考生在学

期末考第二份测验，然后接受定锚测验，以估计两组考生在能力及其他方面的差异。

	(1) 单组设计							
	测验							
	样本组	X	Y					
	P₁	ü	ü					
	(2) 定锚测验随机分组设计							
	测验							
	样本组	X	Y	V				
	P₁	ü		ü				
	P₂		ü	ü				
	(3) 定锚测验不等组设计							
	测验							
	样本组	X	Y	V				
	P₁	ü		ü				
	Q₁		ü	ü				
	(4-1) 分测验预先衔接设计（一个分测验）							
	分测验							
	样本组	X₁	X₂	X₃	Y₁	Y₂		
	P₁	ü	ü	ü	ü			
	P₂	ü	ü	ü		ü		

	P₃	ü	ü	ü			ü		
(4-2)	分测验预先衔接设计（二个分测验）								
	分测验								
	样本组	X₁	X₂	X₃	X₄	Y₁	Y₂	Y₃	Y₄
	P₁	ü	ü	ü	ü	ü	ü		
	P₂	ü	ü	ü	ü	ü		ü	
	P₃	ü	ü	ü	ü		ü	ü	
	P₄	ü	ü	ü	ü	ü		ü	
	P₅	ü	ü	ü	ü		ü		ü
	P₆	ü	ü	ü	ü			ü	ü
(5)	试题预先衔接设计								
	分测验								
	样本组	V₃	W₁	W₂	X₁	X₂	Y₁	Y₂	Z₁
	P₁	ü	ü	ü					
	P₂		ü	ü			ü		
	P₃		ü	ü					ü
	Q₁	ü			ü	ü			
	Q₂				ü	ü		ü	
	Q₃				ü	ü			ü

图一 衔接数据收集设计（数据源：Petersen, Kolen, & Hoover, 1989）

(4-1) 分测验预先衔接设计（一个分测验）：此法旨在使新测验中的某个分测验（也许只是草稿性的试题）能和其他相似内容的完整测验一起实施，然后一起校准后，利用新校准的试题参数估计值来衔接新的测验。例如，如图一之(4-1)所示，让每组考生

接受原测验（共有三份分测验）及加考一份新测验的分测验，故有三种随机方式分派测验，反应数据则可依据任何一种衔接方式来校准参数值。

(4-2)分测验预先衔接设计（二个分测验）：如同前述，如果新测验的分测验有很多，则可以随机方式，安排两种或两种以上的分测验和原测验一起进行施测，再根据结果来校准新测验的参数值。

(5)试题预先衔接设计：此法主要是针对题库发展所需而来。在设计时，唯一需遵循的原则是：不论考生分成多少组，每组至少有一部份试题和其他一组相同。如图一

之(5)所示，测验 \mathbf{M} 和分测验 \mathbf{Y}_2 、 \mathbf{Y}_1 及 \mathbf{Z}_2 以随机方式对三组考生施测；测验 \mathbf{X} 和分

测验 \mathbf{Y}_2 、 \mathbf{Y}_1 及 \mathbf{Z}_1 也以随机方式对三组考生施测，并且将反应数据以同时校准法求得

参数估计值，由于两次校准中都有共同的分测验 \mathbf{Y}_2 ，所以可以将两次校准参数转换到同一量尺上。

上述这些衔接方法及设计，其主要目的是在将两份测验的试题参数估计值，转换到同一量尺上，以便进一步进行测验分数的等化工作。其中，定锚测验设计是最可行的一种等化方法，我们将于下文再继续讨论。

参考书目

1. Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp.508-600). Washington, DC: American Council on Education.
2. Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, 9, 413-415.
3. Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. *Journal of Educational Measurement*, 22, 197-206.
4. Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice*, 7, 29-36.
5. Liou, M. (1990). Effect of scale adjustment on the comparison of item and ability parameters. *Applied Psychological Measurement*, 14, 313-321.
6. Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.
7. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
8. Vale, C. D. (1986). Linking item parameters onto a common scale, *Applied Psychological Measurement*, 10, 333-344.

第十章

試題反應理論的介紹(十) 測驗分數的等化(下) ... (Test score equating)

政大教育系教授 余民宁 着

上文提到定锚测验设计(anchor-test design)是较常用、也较可行的一种等化方法(equating method)，它的主要目的是在利用线性转换的衔接方式(参见上文所列之方法)，将转换所需的常数值(如： α 和 β)加以量化、估算出来，以达到等化的目的。它常常使用下列四种方法来量化常数值：

1. 回归法(regression method)：由于定锚测验是指两组考生(A 或 B)所接受的共同测验，因此，一旦两组在共同测验试题上的试题参数被估计出来后，我们便可套用下列的回归方程式：

$$b_B = \alpha b_A + \beta + \epsilon \quad (\text{公式一})$$

将 α 和 β 值估计出来。其中，公式一中的 ϵ 为回归线的误差项， b_B 和 b_A 为其同试题的两组试题难度参数估计值。所得的回归系数估计值 $\hat{\alpha}$ 和 $\hat{\beta}$ 如下：

$$\hat{\alpha} = r(S_B / S_A) \quad (\text{公式二})$$

$$\hat{\beta} = \bar{b}_B - \hat{\alpha} \bar{b}_A \quad (\text{公式三})$$

其中， r 为 b_B 和 b_A 之间的相关系数， \bar{b}_B 和 \bar{b}_A 为其平均数，而 S_B 和 S_A 为其标准偏差。

若以重复考生为设计的重点时，则回归方程式可以改为

$$\theta_T = \alpha \theta_X + \beta + \epsilon \quad (\text{公式四})$$

其中， θ_T 和 θ_X 为考生在 Y 和 X 测验上的能力估计值。至于 α 和 β 的估计值求法，则和公式二与公式三雷同，但以 θ 估计值代替其试题难度估计值来求解。

使用回归法会有一个问题存在，那就是对称性(symmetry)条件无法满足；也就是说，以 b_A 来预测 b_B 所得的回归系数，并不会和以 b_B 来预测 b_A 所得的回归系数相同。因此，回归法在决定量化的常数值（即 α 和 β ）时，严格说来，并不是一种很适当的作法。

2. 平均数和标准偏差法(mean and sigma method)：由于

$$b_B = \alpha b_A + \beta \quad (\text{公式五})$$

所以，我们可利用简单的代数求得

$$\bar{b}_B = \alpha \bar{b}_A + \beta \quad (\text{公式六})$$

$$S_B = \alpha S_A \quad (\text{公式七})$$

和

$$\alpha = S_B / S_A \quad (\text{公式八})$$

$$\beta = \bar{b}_B - \alpha \bar{b}_A \quad (\text{公式九})$$

同时也可以求得

$$b_A = (b_B - \beta) / \alpha \quad (\text{公式十})$$

一旦 α 和 β 常数值被决定后，在 X 测验上的试题参数估计值，便可顺利的转换到与 Y 测验相同的量尺上，其间的关係如下：

$$b_Y^* = \alpha b_X + \beta \quad (\text{公式十一})$$

$$\alpha_Y^* = \alpha_X / \alpha \quad (\text{公式十二})$$

其中的 b_Y^* 和 α_Y^* 便是转换到 Y 测验量尺的 X 测验之难度和鉴别度值。若使用一个参数型模式的话，则因 $\alpha = 1$ 的缘故，所以上述公式简化成：

$$b_B = b_A + \beta \quad (\text{公式十三})$$

同理，可求得

$$\delta_s - \delta_a + \beta \quad (\text{公式十四})$$

$$\beta = \delta_s - \delta_a \quad (\text{公式十五})$$

由此可知，在 X 测验的难度估计值只要加上共同试题的平均难度值之差，便可转换到 Y 测验的量尺上。

3. 韧性平均数和标准偏差法(robust mean and sigma method): 由于前法未考虑试题参数的估计标准误，因此，Linn, Levine, Hastings, & Wardrop(1981)提出本法，把参数估计值的标准误考虑在内：亦即，把共同

试题 i 的每对估计值 $(\hat{\theta}_s, \hat{b}_s)$ 的较大变异数的倒数考虑在内，具有较大变异数的每对估计值给予较小的加权值(weights)，而具有较小变异数者则予以较大的加权值。其中，难度值的变异数可由其讯息矩阵的对角线元素的倒数而得；对于三个参数型模式而言，其讯息矩阵为一个 3×3 阶的矩阵，而一个参数型模式而言，其讯息矩阵则为一个 1×1 阶的矩阵，亦即只有单一个元素。

本方法的实施步骤可以摘要如下：

- a. 就每对估计值 $(\hat{\theta}_s, \hat{b}_s)$ 来说，加权值 w_i 可以下列方式来决定：

$$w_i = [(\text{最大值 } v(\hat{\theta}_s), v(\hat{b}_s))]^{-1} \quad (\text{公式十六})$$

其中， $v(\hat{\theta}_s)$ 和 $v(\hat{b}_s)$ 是共同试题估计值的变异数。

- b. 求出加权值如下：

$$w'_i = w_i / \sum_{j=1}^K w_j \quad (\text{公式十七})$$

其中， K 为两份测验 X 和 Y 中之共同试题的数目。

- c. 计算出加权后的估计值如下：

$$\hat{b}'_s = w'_i \hat{b}_s \quad (\text{公式十八})$$

$$\hat{b}'_s = w'_i \hat{b}_s \quad (\text{公式十九})$$

- d. 计算出加权后试题参数估计值的平均数和标准偏差。

e. 利用上述所计算出的平均数和标准偏差，来估计 α 和 β 常数值。

上述方法与详细步骤，读者可自行参阅相关文献(Stocking & Lord, 1983; Hambleton & Swaminathan, 1985)。

4. 特征曲线法(characteristic curve method): 由于前两种方法都忽略鉴别度参数在决定量化常数 α 和 β 中所扮演的角色，Haebara(1980)和 Stocking & Lord(1983)乃提出本法，同时考虑难度和鉴别度参数，以补充前法的不足。在实际作法上，先计算两位具有相同能力值 θ_k 的考生，在一个共同试题的两份测验上的真实分数(true scores):

$$\tau_{ik} = \sum_{i=1}^k P(\theta_k, b_{ik}, a_{ik}, c_{ik}) \quad (\text{公式二十})$$

$$\tau_{ik} = \sum_{i=1}^k P(\theta_k, b_{ik}, a_{ik}, c_{ik}) \quad (\text{公式二十一})$$

由于是使用共同试题，下列公式亦会成立：

$$b_{ik} = \alpha b_{ik} + \beta$$

$$a_{ik} = a_{ik} / \alpha$$

$$c_{ik} = c_{ik}$$

所以， α 和 β 常数可经由求出下列 F 函数的极小值而获得：

$$F = \frac{1}{N} \sum_{ik=1}^N (\tau_{ik} - \tau_{ik})^2 \quad (\text{公式二十二})$$

其中， N 为考生人数； F 函数是 α 和 β 所构成的函数，并且是 τ_{ik} 和 τ_{ik} 之间差距的指标。至于计算 α 和 β 常数的算法，是以递归的方式(iterative

method)来进行的，详细过程可以参见 Stacking & Lord(1983)的说明。

在使用定锚测验的设计法中，定锚试题的数目及其特征，对衔接的质量而言，扮演着极为重要的角色。例如，如果所使用的定锚试题对某组考生而言太简单，对另一组考生而言太困难的话，则这两组所获得的试题参数便会显得不稳定，因此使得衔接的质量下降。所以，所使用来作为定锚试题的共同试题，必须具有被两组考生所接受的难度值才能。实证研究结果显示，如果所使用的共同试题均能代表两份即将被衔接的测验内容，则衔接的效果将会是最好。此外，确保这两组考生在能力分配（至少在共同试题）上，具有高度的相似性，也是一件很重要的事。至于定锚试题的数目应该是多少？学者们(Hambleton, Swaminathan, & Rogers, 1991)的建议是：大约是测验试题数的 20% 到 25% 之间。

其他的衔接与等化方法

另一项尚待进一步研究才能证实的方法也可以使用，它的步骤相当简单易行：

1. 把资料看成是由 $(N_x + N_y)$ 位考生接受一份 $(n_x + n_y + n_c)$ 个试题的测验结果，其中的 n_c 是指定锚试题的数目。
2. 把其中 n_y 个试题是 N_x 位考生所没有接受施测的部份，看成是「未答完」(not reached)试题，并予以登录为未答者；同理，把 n_x 个试题是 N_y 位考生所没有接受施测的部份，看成是「未答完」者，并予以登录。
3. 进行试题和能力参数的估计。除了上述各种衔接方法外，试题反应理论也可以应用到下列二方面：(a)两份测验的真实分数之等化；(b)使用特定 θ 值所构成的实得分数分配来进行两份测验的等化。实际的应用过程与衔接步骤，读者可以参阅 Lord(1980) 和 Hambleton & Swaminathan(1985) 的专书说明。底下仅举出两个例子作补充说明，第一个是谈论题库(item bank)发展中的衔接过程的作法，第二个是讨论衔接两份测验的问题。

例子一

假设我们要增加 15 个新试题到一个现成的题库里，我们可以从该现成的题库中，挑选任意 5 个能够与这 15 个新试题之内容和难度值相当的试题，当作是定锚试题，并假设已知其难度值分别为 1.65, 1.20, -0.80, -1.25 和 2.50。

接下来，便是采用方法较单纯易懂的平均数和标准偏差法，来进行量化常数的衔接工作，其步骤如下：

1. 将此 20 题的测验（含 15 个新试题和 5 个定锚试题）给适当的样本（假设为 200 名考生）施测。
2. 选择适用于题库和此 20 题的测验的试题反应模式（假设经过模式与适合度之考验后，一个参数型模式适用于本例）。
3. 计算这 5 个来自题库的定锚试题（假设以 Y 表示）的平均难度值 \bar{b}_y ，其值为 0.66。

4. 接下来，以计算机程序（如：BICAL）来校准这 20 题的测验，在估计的过程中，这 20 题测验的平均难度值设定为零，并计算其中 5 个定锚试题的平均难度值 b_x （假设）为 0.25。

5. 由于共同试题的试题难度值具有下列的线性关系：

$$b_y = b_x + \beta$$

所以 β 值可由 $b_y - b_x$ 计算而得，为 $\beta = 0.66 - 0.25 = 0.41$

（注：在一个参数模式下， $\alpha = 1$ 为已知数，故不予估计）。

6. 然后，这 15 个新试题的试题难度估计值各加上 $(\delta_y - \delta_x) = 0.41$ ，以调整成新的估计值。
7. 同理，这 20 题测验中的 5 个定锚试题的难度估计值，亦各加上 0.41，调整成新的估计值。由于调整后的难度值会与其原本在题库中的已知难度值不同，因此可将调整后的新值与其在题库中的已知值加以平均，作为修正后的共同试题之难度估计值。
8. 至此，这 15 个新试题已和原先现成的题库试题，建立在同一个量尺上了。因此，这些试题便可以正式加入题库，同时，作为定锚试题的 5 个共同试题的难度估计值，也已加以修正过。

上述的计算过程，已摘要在表一里。

表一 把新试题（测验 X ）建立在题库试题（测验 Y ）量尺上的衔接过程

试题	测验 X 的难度 b_x	测验 Y 的难度 (共 b_y 同试题)	量化后测验 X 的 难度 $b_x + (\delta_y - \delta_x)$	量化后测验 X 的难度(修 正值)
1	1.29	1.65	1.70	1.67
2	0.75	1.20	1.16	1.18
3	-1.24	-0.80	-0.83	-0.82
4	-1.72	-1.25	-1.31	-1.28
5	2.17	2.50	2.58	2.54
6	0.85		1.26	1.26
7	-1.88		-1.47	-1.47

8	-2.02		-1.61	-1.61
9	0.19		0.60	0.60
10	0.22		0.63	0.63
11	-1.86		-1.45	-1.45
12	-1.32		-0.91	-0.91
13	-1.10		-0.69	-0.69
14	0.74		1.15	1.15
15	0.61		1.02	1.02
16	0.50		0.91	0.91
17	-0.80		-0.39	-0.39
18	1.70		2.11	2.11
19	1.37		1.78	1.78
20	1.55		1.96	1.96
	$\bar{d}_x = 0.25$	$\bar{d}_y = 0.66$	$\bar{d}_y - \bar{d}_x = 0.41$	

注：共同试题以粗写体字印刷

例子二

假设我们要衔接两份不同试题的能力测验，各具有 15 个试题。我们可以设计一些具有代表性内容的共同试题，其内容性质约略与这两份测验（分别以 X 和 Y 来表示）相当，假设这份定锚测验共有 6 个试题，我们便可放入这两份测验里，共同进行校准其试题参数值。

假设所选用的三个参数模式（其中的 C 值为 0.2）适合本研究，我们便可使用平均数和标准偏差法来进行衔接这两份测验，其步骤如下：

1. 计算 X 和 Y 测验中共同试题的难度估计值的平均数和标准偏差（可利用 LOGIST 或 BILOG 等计算机程代为计算）。
2. 决定常数 α 和 β 的估计值。
3. 将 X 测验中的难度估计值乘上 α ，再加上 β ，以转换到 Y 测验的量尺上。
4. 把共同试题的难度估计值加以平均。

5. 将 Y 测验中的鉴别度估计值除以 α ，以转换到 Y 测验的量尺上。
6. 把共同试题的鉴别度估计值加以平均。

上述计算过程如表二和表三所示，如此一来， X 测验上的难度和鉴别度参数估计值便已和 Y 测验上的试题参数估计值，建立在同一量尺上了。

表二 量化常数值的决定与 X 和 Y 测验中已量化后的难度值

试题	测验 Y 的难度	测验 X 的难度	量化后所有试题的难度
1	1.20	1.20	1.20
2	1.75	2.10	1.75
3	-0.80	2.75	-0.80
4	-1.28	-1.40	-1.28
5	1.35	-1.65	1.35
6	1.40	0.60	1.40
7	1.20	1.81	1.20
8	0.50	2.20	0.50
9	0.72	2.70	0.72
10	-1.95	1.86	-1.95
11	-2.20	-0.90	-2.20
12	2.40	-1.10	2.40
13	1.80	-2.30	1.80
14	1.45	0.58	1.45
15	0.80	0.92	0.80
16	1.10	0.88	1.03
17	1.85	1.92	1.83
18	2.30	2.10	2.36

19	-1.50	2.52	-1.51
20	-1.80	1.60	-1.78
21	0.40	-1.20	0.40
22			1.54
23			1.91
24			2.38
25			1.59
26			-1.04
27			-1.23
28			-2.37
29			0.37
30			0.69
31			0.66
32			1.64
33			1.82
34			2.21
35			1.34
36			-1.32
	$b_Y = 0.39$	$b_X = 0.60$	$\alpha = 0.95$
	$S_Y = 1.56$	$S_X = 1.65$	$\beta = -0.18$

注：共同试题以粗写体字印刷

表三 \mathbf{X} 和 \mathbf{Y} 测验的鉴别度值

试题	测验 \mathbf{Y} 的鉴别度	测验 \mathbf{X} 的鉴别度	量化后所有试题的鉴
----	----------------------	----------------------	-----------

			別度
1	1. 20	0. 90	1. 20
2	1. 21	1. 15	1. 21
3	0. 90	1. 86	0. 90
4	0. 72	0. 55	0. 72
5	1. 25	0. 40	1. 25
6	1. 40	0. 65	1. 40
7	1. 12	1. 60	1. 12
8	0. 75	1. 85	0. 75
9	0. 92	1. 90	0. 92
10	0. 62	1. 62	0. 62
11	0. 52	0. 81	0. 52
12	1. 98	0. 62	1. 98
13	1. 90	0. 40	1. 90
14	1. 62	0. 64	1. 62
15	1. 01	0. 80	1. 01
16	0. 95	0. 75	0. 95
17	1. 23	1. 23	1. 22
18	2. 00	1. 55	1. 98
19	0. 68	1. 72	0. 63
20	0. 45	1. 12	0. 44
21	0. 70	0. 42	0. 69
22			1. 68

23			1.95
24			2.00
25			1.70
26			0.85
27			0.65
28			0.42
29			0.67
30			0.84
31			0.79
32			1.29
33			1.63
34			1.81
35			1.18
36			0.44
			$\alpha = 0.95$

注：共同试题以粗写体字印刷

我们也可以利用 α 和 β 等常数值，来衔接参与 X 和 Y 测验的考生能力估计值。

由于参与这两份测验的能力参数衔接公式为：

$$\theta_t = \alpha\theta_x + \beta = 0.95\theta_x - 0.18$$

所以，在估计过程中，接受 X 测验的这组考生能力值是设定为零，因此，它相当于接受 Y 测验这组考生的能力值：

$$\bar{\theta}_t = 0.95(0) - 0.18 = -0.18$$

这意谓着接受 X 和 Y 测验的两组考生的平均能力差为 -0.18；亦即，接受 X 测验这组的考生平均能力，比接受 Y 测验这组的考生平均能力，要低 0.18 个单位估计值。这项涵意在学术研究和课程评鉴上，具有重大的启示。

参考书目

1. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer.
2. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
3. Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

本文转载自研习信息 10 卷 (3 期), 11-16 页

第十一章

試題反應理論的介紹(十一)題庫的建立 ... (Item Banking)

政大教育系教授 余民宁 着

题库 (item bank 或 item pool) 不光只是一堆试题的集合体而已，而是一堆经过校准(calibration)、分析、归类、与评鉴后，贮存起来的测验试题组合体。Millman & Arter(1984)便将题库界定为一群使用方便的试题汇编；他们的意思是说该群试题可资应用于各种测验场合的数量非常庞大，并且都是经过分析、编码、与结构分类处理后的试题，并且有逐渐走向计算机化的趋势。

题库（尤其是以试题反应模式参数估计值所建立起来的试题）具有下列的优势：

1. 可使测验编制者（也许是教师或专业机构）随心所欲地编制能够符合各种目标的测验。
2. 可使测验编制者就题库的范围内，编制出每个目标都有适当题数的试题来测量到它的测验。
3. 如果题库能够包含内容有效且编题技巧纯熟的试题在内的话，则测验质量通常会比测验编制者自己编的测验质量还好。

由此可见，题库具有改进测验质量的潜能，在可预期的将来，它对测验编制者的重要性将日益增加，同时对节省编制测验所花的时间，亦将无可限量(Hambleton & Swaminathan, 1985, PP. 255-256)。

建立题库的步骤

题库的建立，当然是依据课程标准、教材大纲、双向细目表的编写而成，它的建立过程，可以分成下列十一个步骤（许择基、刘长萱，民 81）：

1. 试题的编写与修订：首先，仿照传统编制测验的原则，撰写大量的试题，并邀请学科专家（如任课教师）和测验专家就试题进行形式审查，看看是否能符合内容效度的要求，否则加以修改或删减。
2. 选题预试：放在题库里的试题，都必须是建立在同一量尺上的才行，否则试题间无法比较或延用。因此，选择适当的试题和考生样本，是一项很重要的步骤，幸好前二文所谈的定锚测验设计(anchor test design)可以提供本步骤的参考(Vale, 1986)：
 - a. 定锚试题的数目：若使用同时校准法，则至少必须使用两题定锚试题；若使用等组法，则可以不用定锚试题。一般而言，若将来所编制的测验，具有 60 到 80 个试题的话，则在题库建立过程中，必须使用 15 至 20 个定锚试题才够。
 - b. 每个定锚组别至少要包含 30 位考生。

至于考生样本数要多大才算足够？大致上可以这么说：若使用二个参数或三个参数对数型模式来进行校准时，则至少必须使用 1000 位考生；若使用一个参数对数型模式的话，则可以减少到 500 位考生便行。样本的能力范围，最好是呈常态分配。

3. 试题的校准与衔接：选择适当的反应模式来分析数据，必须考虑试题的性质。就选择题而言，当然是以三个参数对数型模式最适合。决定好适当模式后，便可采用适当的计算机程序（如：BILOG3 或 LOGIST5 等），以进行试题参数与考生能力参数的估计与适合度分析，统称为校准(calibration)。经过校准后的试题，必须能够通过适合度的考验者，方可被保留在题库里，因为它们可以被适当的反应模式所解释。如果在校准时，所使用的是不同的考生样本，则在将试题放入题库之前，还必须做试题衔接的工作（请参考前二文的说明），如此才能将所有的试题参数都建立在同一个量尺上。
4. 更新题库：理想的题库特色是，包含题数相当充份的试题、试题具有内容效度、鉴别度不低于 0.8 以下、难度分布均匀、猜测度愈小愈好等。并且由于试题被选入不同的测验里，和不同的试题出现在同一份试卷中，在施测时会产生不同的背景影响(context effect)。因此，当题库里的试题被选用之后，都必须有详实的施测记录，甚至必须再重新校准一次，以确定该试题参数的真正适合度。如此可以确保题库之素质能够不断地更新，也可以保持题库之安全，避免沦为考古题而被众多考生熟悉，因而丧失题库的功能。另外，也可以视测验目的、使用题库的目的、和学科的性质，于每次施测前，重新组合与排列题库中的试题，以方便未来的使用。
5. 测验编辑：如果题库的素质很高，则从题库中抽取试题来编制一份测验，便会很容易。编辑测验的方式很多，最主要是看测验目的而定。往往是由专家先将试题按学科、单元、属性、和概念等，先行予以计算机编码，再按其他考虑事项（如：试题参数值、讯息函数值、估计标准误等），撰写在计算机程序里，以便编辑时输入几个关键词，就可获得想要的测验。
因为题库的内容庞大，几乎不太可能用人工选题的方式，来编印试卷。通常都是仰赖计算机的帮助，因此在打印试卷上，也有几种方法可供参考：
 - a. 分层随机抽样选取试题：按教材内容来分，将题库予以分成几个层次，然后就每个层次中随机抽取适当的题数，以作为打印试卷的内容。这

- 种作法，无法保证被选出的试题质量就一定是最好的。
- b. 依试题参数值随机抽样：测验编制者可依据教材内容，决定具有所欲的难度、鉴别度、和猜测度的试题参数范围，及拟使用的题数，再由计算机自合格的试题中随机抽样，以编成一份试卷。这种作法的最大优点便是，免除人为的偏见，并确保试题具有一定的质量。
 - c. 由试题讯息量来选取试题：首先，由测验编制者决定理想的目标讯息曲线(**target information curve**)（读者可以参阅前面「讯息函数」与「测验编制」二文），然后自己校准的试题中，选取讯息量能够填满此一曲线的候选试题，可中途更换较佳的候选试题，每选出试题便计算其讯息量是否已接近理想的曲线，若否，则一直继续这种选题过程，直到理想的目标讯息曲线被填满为止。
 - d. 由测验编制者主观选题：测验编制者依据试题的特性和统计分析的资料，再由本身的专业判断，以便决定选取何种试题。
6. 评估测验质量：对于新编制的测验，可用试题反应理论所适用的计算机程序（如：**BILOG3** 和 **LOGIST5**）来预测其特性。例如，计算机程序可利用所选取试题的难度、鉴别度、和猜测度的估计值，来计算出试题参数估计值的平均值、信度估计值、测验讯息期望值和平均值、和各种不同长度下的预期测验讯息量等数据，以便让测验编制者来判断所编测验的优劣。如果所编的测验不符理想，则可以依据前述步骤来重编。
7. 测验是否达预期的水平：根据第六步骤的资料，来判断所编的测验是否有达到预期的水平：如果达到，则进行第八步骤；如果尚未达成，则回到第四步骤，重新更新试题再来。
8. 执行考试：如果前个步骤显示测验质量不错，则可对考生进行施测。当然，施测应有的指导语、测验情境的安排与布置、和其他会影响考试的注意事项等，都应该事前的准备与策划。
9. 评分：在经过考试后的学生作答数据，可再被拿来行进行试题校准，此时，学生的考试成绩，可用下列二种方法之一来加以评分：
- a. 直接以学生的能力估计值 $\hat{\theta}$ 来代表学生的能力。唯这种作法，比较不容易被大众所了解，因此解释起来，颇费周章。
 - b. 以真实分数(**true score**)来表示学生的能力。亦即将每位考生在每个试题上的答对机率，加总起来的和，即是他的真实分数。真实分数的值域将分布于全部试题的猜测度之和与试题总题数 (n) 之间。唯这种作法，仍有其解释上的不便处，因此，可将真实分数除以试题总题数，以转换成正确答对试题的百分比分数，此分数则与一般学校惯用的百分制计分方式的意义相同：愈接近百分之百，表示其能力愈高；反之，愈接近零，则表示其能力愈低。
10. 决策：此步骤旨在应用上述评分与试题评鉴的结果，作为甄选学生，诊断命题技巧，与改进教学的参考。
11. 研究与评鉴：题库的应用，不仅是用于编制新测验，以节省人力、物力、和时间，并可透过每次考试完毕后，针对试题与考生能力参数进行校准，以评鉴试题质量的好坏、试题内容有否偏差（如：有利于某种族群的考生，而不利于另一种族群的考生）、以及诊断学生的作答数据有否不寻常、或找出学习有误差

的部份等，这种不断研究与评鉴的过程，正是题库所提供的特色。

发展题库的适当时机

已知上述建立题库的步骤后，什么情况下，才需要去建立并运用题库？Millman & Arter(1984)建议在至少满足下列条件之一的情况下，才需要着手建立题库，并充份发挥题库的价值。

1. 现存测验无法广被接受，并且客观环境需求编制属于自己的测验时。
2. 经常需要进行测验时。
3. 需求具有多份复本测验时。
4. 实施个别化适性测验(individualized adaptive testing)时。
5. 许多测验使用者愿一致建立满足自己所需的题库时。
6. 已具备题库系统，如：计算机设备和可用之计算机软件时。

建立题库所面临之重大课题

1. 题库应该包含多少试题？

基本上而言，题库内的试题当然是愈多愈好，但是应该考虑所加入题库的试题，是否具有内容效度和统计质量应达成的标准，以及考虑测验的目的何在？

Prosser(1974)建议每个概念至少要包含 10 个试题，每一单元课程内容至少要包含 50 题。Reckase(1981)则建议一百至二百个难度均匀分布，且具有合理的鉴别度的试题，便可适用在计算机化适性测验(computerized adaptive testing)里。另外，测验的目的如果是在对课程作一整体的评估，则不需针对每项学习细节编制太多试题；如果仅在作学习诊断，则许多学习细节部份，仍需要编制许多试题去测量它们。

2. 题库系统应该如何分类？

常见的分类系统是依据内容来分类，它有两项作法：一是依主题或教学目标来检索试题，另一是采关键词方式来检索试题。一般而言，采关键词方式比较富弹性，可以同时适用于目标、内容、年级、及思考历程等；但是依主题或教学目标检索方式，则比较可以显现知识结构的层级分明。当课程修订时，采用关键词系统者修订比较容易、迅速；但如果计算机无法处理多重关键词时，或分类系统本身就具有明确的界定（如：生物学中的种、属、科、目等层次的分类）时，则采用固定的分类方法就比较适合。专家们的建议，任何题库系统都应兼具这两种编码检索的方法。

3. 题库内试题是否必需具备量尺化的参数？

所谓量尺化的试题参数，是指将试题参数（如：难度值）经过校准后，都换算成同一量尺单位的指针。这种参数，正是试题反应理论具有试题参数不变性使然，也是古典测验理论所采用的试题参数指针（但容易受样本影响）所无法媲美的。因此，如果能对大样本进行施测，则试题参数的量尺化就非常必要；但如果仅能对教师个别班级施测的话，则试题参数的量尺化问题便可予以忽略，因为不仅是不可能，而且也没有必要。至于，如果学校仍想运用试题参数的量尺化过程，来建立起属于自己学校适用的题库的话，则本文作者的建议：不妨实行几个学校联合命题的方式，以力求获取大样本（如：大于 1000 人以上），来建立起量尺化的参数试题。

4. 题库是否可以公开？

如果题库可以公开，让任课老师任意取用，则有人担心：教师是否从此就仅教题库内容，而使教学活动更形窄化。这点忧虑也是必然的现象。不过，有个观点必须厘清：由于建立一个量尺化的题库，不是一件容易的事，它必须投入大量的人力、物力、时间、和金钱，才能有所斩获，因此只要题库够大（理论上可以达到无穷大），教学是否会因此窄化的问题，倒可以不必担心，因为教师无法做到仅教题库，而不教正常的课程内容；但是，如果题库不够大的话，公开题库必然导致窄化教学活动，因此，是否要完全公开，则有待进一步的商榷。但是，基本上公开少数的样本试题，以让教师和考生明了运用题库的评量方式和重点，则是正确、并且有其必要的作法。

5. 题库是否安全？

题库的建立，固然可以使日后的测验编制更加容易，也可以使评量问题更轻松地获得解决。但是，题库的重复使用，是否会妨碍到试题的安全性（如：雷同或考古试题的出现）？这点忧虑，也许在题库少时，会有此顾虑，但在题库大时，这层顾虑也许就可以不必有。另外，随时更新题库的内容，确保试题的内容效度和统计质量，也是保障题库安全的一项作法。

参考书目

1. 许择基、刘长萱（民 81）。试题作答理论简介。台北：中国行为科学社。
2. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.
3. Millman, J., & Arter, J. A. (1984). Issues in item banking. Journal of Educational Measurement, 21, 315-330.
4. Prosser, F. (1974). Item banking. In G. Lippey (Ed.), Computer-assisted test construction(pp. 29-66). Englewood Cliffs, NJ: Educational Technology.
5. Reckase, M. D. (1981). Tailored testing, measurement problems and latent trait theory. Paper presented at the annual meeting of the National Council for Measurement in Education, Los Angeles.
6. Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, 333-344.

本文转载自研习信息 10 卷 (4 期) , 9-13 页

第十二章

試題反應理論的介紹(十二)
.... 電腦化適性測驗 ...
(Computerized Adaptive Testing)

政大教育系教授 余民宁 着

前文谈到测验题库的建立，建立题库的另一个优点，就是为计算机化适性测验

(computerized adaptive testing 简写成 CAT) 作准备，不仅可以节省施测时间，更可以达到精确估计考生能力或某种潜在特质的目的。

从前面几篇文章里，我们也大致可知道：当测验的难度能够适合考生的能力程度时，这时测验所测量到的考生能力最为精确。因此，我们可以知道任何一次施测的结果，都无法针对每位考生提供最精确的能力测量，因为该测验的难度无法适合每位考生能力程度的需求。最理想的施测状况是：能够针对每位考生不同的能力程度，来提供适合个别情况的测验方式，这也就是计算机化适性测验所欲探讨的内容。

最早应用适性测验（即因材施测式的测验方式）的例子，是 1908 年 Binet 所作的智力测验的研究(Weiss, 1985)。后来一度中断好久，直到 1960 年代末期，才由在教育测验服务社(Educational Testing Service)的 F. Lord 从事较为完整的通盘研究（参见 Lord(1980)的作品）由于 Lord 感觉到，对于低能力与高能力的考生而言，固定长度的测验无法有效的满足这些考生能力估计的需求，因此才极力投入适性化测验的研究。Lord 认为如果被挑选用来施测的试题都能针对每位考生能力提供最大的参考讯息的话，则缩短测验的长度（即减少施测的题数），应该不会降低对每位考生能力的精确测量。理论上而言，每位考生所接受的施测试题，应该都是不同的试题组，因此，适性测验的实施是可能的。

但是要实施适性测验，也唯有在计算机诞生发明后，才有可能施行。计算机科技的发达，日新月异，它的超大容量可以贮存测验讯息（如：测验试题及其特征指针）、编制、施测、和记录测验分数，因此使得推行适性测验变得愈来愈可行(Bunderson, Inouye, & Olsen, 1989; Wainer, 1990)。在 1960 年代末，美国陆军总署、人事管理局、及其他联邦机构，均大力支持赞助有关适性测验的研究，举办特殊的专题研讨会，并且有数百篇的有关研究论文发表出来，后来都被收集出版成册(例如: Wainer, 1990; Weiss, 1983)。

在计算机化适性测验(CAT)里，呈现给考生的试题顺序，是依据考生在前一个试题上的表现好坏来作决定的。根据考生先前的表现好坏，下一个要呈现给考生作答的试题，便是对考生能力估计精确性最有贡献的最大讯息量的试题。如此一来，测验的长度便可以缩短，并且也不会牺牲任何的测量精确性；因为对于高能力的学生，可以不必给他相当容易的试题作答，对于低能力的学生，也可以不必给他极度困难的试题作答，因为这些试题对他们的能力水平的估计而言，只能提供极为有限或丝毫没有帮助的讯息。因此，实施计算机化适性测验，不仅可以做到因材施测的精确估计考生能力的地步，也可以节省许多施测时间和成本，可说是至少一举两得。

在开始进行计算机化适性测验之时，先由计算机终端机随机呈现一组测验试题（也许是两题或三题），在考生作出反应之后，计算机便根据这些反应资料，估计出考生的初步能力估计值(initial ability estimate)；然后，计算机会根据这些初步能力估计值，从现有的题库（储存在计算机的内部）中挑选出最能对能力水平的估计发挥最大贡献力量的试题（通常这些试题的讯息量也是最大），再呈现这些试题给考生作答；这种施测过程一直继续下去，直到事先预定的施测题数已测完，或某种预定的能力估计值的测量精确性（即标准误）已获得为止。

试题反应理论的新天地

我们已知难易适中的试题，对估计考生能力的精确性最为有效。而通常的一份测验卷试题难度，很难满足或适合每位考生的能力水平，因此要能做到试题难度随考生能力不同（即个别差异）而调整的测验方式，唯有实行适性测验。而最适合在适性测验中作应用的，便是试题反应理论(IRT)。由于试题反应理论中的试题反应模式，可以

获得不受不同施测试题影响的能力估计值（即具有试题独立（参见本系列本章之二和之三）的估计特性），也就是说不同的考生考不同的试题，只要试题性质相同，不同能力考生的能力估计值可以被精确的估计出来，因此可以互相比较。事实上，也唯有试题反应理论才适合应用在适性测验里。

在应用试题反应理论到实际的测量情境时，必须先满足该测验只具有单一主要的因素的基本假设（参见本系列文章之二），这个基本假设在目前所使用的适性测验里，一般都能够获得满足。目前，最适合应用到适性测验上的试题反应模式，是三个参数对数型模式（即 3PL）(Green, Book, Humphreys, Linn, & Reckase, 1984; Weiss, 1983)，最主要的原因是它比一个与二个参数对数型模式，更适合用在选择题的试题数据上。

在适性测验里，试题讯息函数也扮演着很重要的角色（参见本系列文章之七）。其中，能对测量精确性发挥最大贡献力量的试题，会被优先挑选做施测的试题，呈现给考生施测。一般而言，能让考生大约有 50% 或 60% 答对机率的试题，通常都是属于能够提供最大讯息量的试题。

适性测验的基本方法

以试题反应理论为架构下的适性测验，有个基本目的，那就是要撮合测验试题的难度和待测量的考生能力水平。为了达成这项目，我们必须拥有已知每个试题特征的庞大试题库，以便从中挑选出适当的试题(Millman & Arter, 1984)。根据 Lord(1980)的看法，我们必须设计计算机程序，以便完成下列的目的，达到适性测验的目标：

1. 根据考生先前的反应表现，预测他在尚未接受测验的试题上的种种可能反应。
2. 根据上述的理解，有效地挑选试题，接着呈现给考生作答。
3. 最后在测验完毕时，能够计分，以分数来表示考生能力的大小。

适性测验的基本方法，包括下列六个步骤。兹简述各个步骤如下：

1. 试题反应模式的挑选：针对不同类型资料和研究问题的了解，挑选适当的一个、二个、或三个参数对数型模式，作为进行适性测验的最基本模式根据。当然，三个参数对数型模式是被最常选用的模式。
2. 题库的准备：参考本系列文章之十一的说明。
3. 测验的起点：应该先考那一个试题，是适性测验所需面临的一件重要抉择问题。从理论上来看，试题的难度必须要能够配合考生的能力水平。但是，除非我们已知考生过去的表现好坏，否则无法在施测之前就知道考生的能力。所以，常用的起点方法有：(1)自难度适中的试题中随机抽取一个试题；(2)完全随机抽取一个试题；(3)先调查学生的背景，然后再决定出那一类的试题。Lord(1977)认为，只要测验的题数不少于 25 题的话，以那一个试题做为起点的影响不大。
4. 选择方式：使用试题反应理论作为适性测验的理论基础，必须有根据某种理论建立的题库存在，以方便经过校准过后的试题参数特征，也都能储存在题库里。校准时所选用的模式不同，也会影响到计分方法的选择和能力的估计。一般而言，常用的试题选择的方法有三种：(1)挑选能对考生能力估计提供最大讯息量的试题，为了避免同样的试题一再地被重复选用，Green 等人(1984)建议可从一堆产生最大讯息量的试题中，随机抽取一个试题来进行就可以。(2)利用贝氏试题选择法，将考生能力分配看成是某种事先分配(prior

distribution), 计算考生答对或答错未用到的试题之后变异数, 再挑选能够使这种考生产能事后分配之变异数为最小的试题, 作为施测的试题。使用贝氏的选题方法, 颇受事前分配之假设的影响很大, 但是只要施测的试题很多的话, 这种影响是可以被排除的。(3)挑选难度最接近考生产现阶段能力估计之试题。

5. 计分方法: 其实也就是学生的能力估计方法, 唯一不同的是, 在适性测验里, 考生每答对一个试题, 就得重新估计一次考生产能估计值。最大近似值估计法和贝氏估计法是适性测验里常用的两种能力估计方法。最大近似值估计法的估计效能很好, 但遇到题数少或估计值无法收敛时, 都会产生很大的问题; 贝氏估计法虽能克服这些困难, 但对事前分配的假设如果不当的话, 却会产生有所偏差的能力估计值。
6. 终止的标准: 终止适性测验的方法, 和前述的选题与计分方法间有很密切的关联。若以试题最大讯息量作为选题标准的话, 只要累积已测过之试题的讯息量, 到达某种事先预定的标准后, 便可终止。若以贝氏估计法来选题的话, 则可以估计能力之变异数小到某个预定的标准时, 便可终止施测。此外, 如果前述两种标准均很慢才达到的话, 也可以预设施测试题的上限, 只要题数一测完, 即使尚未达到预定的标准, 也可以终止施测, 以避免漫无止境地继续下去, 浪费考生的许多宝贵时间。

适性测验的例子

假设从一已知的题库(如: 表一所示)中, 进行适性测验(事实上的题库试题应有数百题, 在此所列举者, 仅作为例子说明用), 则下列的步骤是计算机化适性测验中会出现的事件:

表一 假想的题库试题

试题	<i>b</i>	<i>a</i>	<i>c</i>
1	0.09	1.11	0.22
2	0.47	1.21	0.24
3	-0.55	1.78	0.22
4	1.01	1.39	0.08
5	-1.88	1.22	0.07
6	-0.82	1.52	0.09
7	1.77	1.49	0.02
8	1.92	0.71	0.19
9	0.69	1.41	0.13
10	-0.28	0.98	0.01
11	1.47	1.59	0.04
12	0.23	0.72	0.02
13	1.21	0.58	0.17

1. 假设先挑选试题 3; 因为它具有平均难度值和最高的鉴别度值。又假设某考生答对, 但此时的最大近似值估计法无法进行能力估计, 必须等到至少有一题答

- 对和一题答错才行（全错或全对的得分，会导致 -10 和 $+10$ 的能力估计值）。
2. 假设其次挑选试题 12，因为它比前一个试题较难。又假设该考生答对。至此，最大近似值估计法仍无法进行能力估计。
 3. 其次选中试题 7，因为它比前二题较难。假设该考生答错本题。该考生在三个试题上的反应组型为 $(1, 1, 0)$ ，利用这三个试题的已知特征和最大近似值估计法，估计出该考生能力估计值为 $\hat{\theta} = 1.03$ ；这三个试题的测验讯息量为 $I(\hat{\theta}) = 0.97$ ，估计标准误为 $SE(\hat{\theta}) = 1.02$ ，如表二所示。

表二 每一阶段计算机化适性测验后的能力估计值和估计标准误

阶段	试题编号	试题反应	$\hat{\theta}$	$I(\hat{\theta})$	$SE(\hat{\theta})$
1	3	1	—	—	—
2	12	1	—	—	—
3	7	0	1.03	0.97	1.02
4	4	1	1.46	2.35	0.65
5	11	0	1.13	3.55	0.55
6	9	1	1.24	4.61	0.47
7	2	1	1.29	5.05	0.45
8	1	1	1.31	5.27	0.44
9	8	0	1.25	5.47	0.43

4. 接着，计算出当 $\theta = 1.03$ 时，题库中剩余试题所提供的讯息量，如表三所示。由于试题 4 在 $\theta = 1.03$ 时所提供的讯息量最大，所以，它是下一个被挑选中的试题。假设该考生答对本题，接着估计出反应组型为 $(1, 1, 0, 1)$ 时新能力估计值 $\hat{\theta} = 1.46$ ，估计标准误 $SE(\hat{\theta}) = 0.65$ ，如表二所示。

表三 每一个计算机化适性测验阶段中剩余试题所提供的讯息量

阶段	$\hat{\theta}$	试题所提供之讯息量												
		1	2	3	4	5	6	7	8	9	10	11	12	13
4	1.03	.034	.547	—	1.192	.010	.051	—	.143	1.008	.251	1.101	—	.166

5	1.46	.179	.319	—	—	.004	.017	—	.205	.579	.136	1.683	—	.175
6	1.13	.292	.494	—	—	.008	.039	—	.159	.917	.219	—	—	.170
7	1.24	.249	.433	—	—	.006	.029	—	.175	—	.187	—	—	.173
8	1.29	.232	—	—	—	.006	.026	—	.182	—	.175	—	—	.174
9	1.31	—	—	—	—	.005	.024	—	.186	—	.168	—	—	.174
10	1.25	—	—	—	—	.006	.028	—	—	—	.184	—	—	.173

5. 接下来，计算出 $\theta = 1.46$ 时，剩余试题所提供的讯息量。然后，挑选出最大讯息量的试题。之后，再施测、重新估计能力、计算剩余试题所提供之讯息量、再挑选最大讯息量的试题，如此继续下去（如表二所示，接下来被选中的试题，依序为试题 11、9、2、1、和最后的试题 8），一直到考生能力估计值的估计标准误之递减值小于事先预定的标准（如：小于 .01）。从表二可知，第九个阶段施测试题 8 之后，它从第八个阶段所递减的标准误值为 .01，因此，施测的过程到此为止。此时，该考生的能力估计值为 $\theta = 1.25$ 。这个估计值便是我们从题库中挑选 9 个试题进行适性测验后，所精确估计出该考生的能力水平。

由本例可知，实施适性测验将具有下列的几项优点：

1. 加强测验的安全性；
2. 依据需求来进行施测；
3. 无需使用答案纸；
4. 适合每位考生的作答速度；
5. 立即的计分和报告成绩；
6. 降低某些考生的考试挫折感；
7. 加强施测的标准化过程；
8. 容易从题库中找出并删除不良的试题；
9. 对于试题类型的选择更具弹性；
10. 减低监试的时间。

参考书目

1. Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 367-407). New York: Macmillan.
2. Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. Journal of Educational Measurement, 21, 347-360.

3. Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.
4. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement*, 21, 315-330.
6. Wainer, H. et al. (Eds.). (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.
7. Weiss, D. J. (Ed.). (1983). New horizons in testing. New York: Academic Press.
8. Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 774-789.

本文转载自研习信息 10 卷 (5 期) , 5-9 页

第十三章

試題反應理論的介紹(十三) 試題偏差的診斷 ... (The detection of item bias)

政大教育系教授 余民宁 着

社会大众对心理测验或教育测量有个相当迫切的关注，那就是测验的公平性(test fairness)问题。例如，我们的大学联考试题对少数族群（如：偏远地区、离岛、或少数民族的学生）的考生而言，都很公平吗？我国的高普考试试题对性别不同的男女考生而言，也都很公平吗？这些类似问题的答案，也许都不是。由于编制测验试题的专家，受到自己本身的专业素养、国学程度、文化认知、甚至主观偏见等限制和影响，以致所编制出的试题有时只会有利于某些族群的考生，而不利于另一些族群的考生，这种现象和问题，便是本文所要探讨的试题偏差(item bias)的问题。虽然，在古典测验理论里也谈试题偏差的诊断和补救，但试题反应理论对此问题所提出的理论基础和考验架构，却是相当完整、周延、和严谨的。

传统上对诊断试题偏差的作法是：收集所关怀的少数族群(minority)在测验试题上的表现好坏数据，以及多数族群(majority)的表现数据，再比较其差异，以作为判断试题有否偏差的实征证据(empirical evidence)。其实，表现有差异存在的实征证据是结论说试题有偏差的必要条件，而非充分条件；也就是说，这种结论已超过数据所能推论的范围。为了区别实征证据与结论间的不同，学者们往往使用「不同的试题运作功能」(differential item functioning, 简写成 DIF)一词来取代涵意不明确的「偏差」(bias)概念，以用来描述实征证据背后所涵盖的偏差涵意(Berk, 1982)。

即使对什么样的 DIF 的定义才较适当？也有很多争辩存在。目前有个关于测验公平性问题的看法认为：「在某个试题上，如果多数族群和少数族群的平均表现有所不同的话，该试题便显示出具有 DIF 的现象。」其实，这种看法也有个缺失，那就是未考虑其他影响变项的可能性，如：原本这两个族群的能力就有所不同，因此才导致他

们在某个试题（或某份测验）上表现不同(Lord, 1980)。

目前，比较被心理计量学者所接受的 DIF 的定义为：「来自不同族群，但能力相同的个人，如果在答对某个试题上的机率有所不同的话，则该试题便显现出 DIF 的现象。」有了这项定义，试题反应理论(IRT)很自然的提供一个研究 DIF 的架构，因为试题特征函数正可以说明答对某个试题的机率，是与受试者的潜在能力和试题的潜在特征有某种关联存在。因此，DIF 的定义可以被写成下列的操作型定义：「某个试题特征函数如果对不同的族群而言都不相同的话，则该试题便显现出 DIF；反之，如果跨越不同族群的试题特征函数都相同的话，则该试题便不具有 DIF。」本文即谈论试题反应理论对诊断试题偏差（或说试题 DIF）的各种方法，并举例说明它的用法。

诊断 DIF 的 IRT 方法

根据上述的定义，我们只要比较两个或多个族群在某个试题特征函数上的差异，就可以判别该试题是否具有 DIF 存在。试题反应理论常用来诊断试题偏差的方法有三种：一为比较试题特征曲线的参数；另一为比较介于试题特征曲线间的面积；最后一种为比较反应模式与数据间的适合度。兹分别描述如下：

一、比较试题特征曲线的参数

如果两个试题特征函数的参数值相同的话，则该试题特征曲线在线所有点的功能会相同，答对该试题的正确机率值也会一样。因此，试题特征函数的参数均相等的虚无假设，可以表示如下：

$$H_0: \alpha_1 = \alpha_2; \beta_1 = \beta_2; \gamma_1 = \gamma_2$$

足标表示不同族群的参数估计值。如果我们能够拒绝某个试题的虚无假设，则显示该试题具有 DIF 的现象。

这种诊断的方法，需要用到参数估计值的变异数一共变量矩阵（或讯息函数矩阵），其诊断的步骤如下：

1. 选取一个适当的试题反应模式。
2. 分别估计不同族群考生的能力及试题参数。
3. 经由衔接的过程，将参数值建立在共同的量尺上。
4. 以矩阵表示试题参数所组成的向量，例如： $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]$ ，并计算其讯息矩阵或变异数一共变量矩阵。
5. 计算虚无假设的统计考验值如下：

$$\chi^2 = (\boldsymbol{X}_1 - \boldsymbol{X}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_1 - \boldsymbol{X}_2)$$

其中， $\boldsymbol{\Sigma}$ 表示是参数估计值之差值的变异数一共变量矩阵。此 χ^2 统计值将成为 P 个自由度的卡方分配， P 为我们所选用的试题反应模式的参数个数；例如，选用三个参数对模型模式时， P 为 3；选用二个参数对模型模式时， P 为 2。

6. 选定临界点（如 $\alpha = .05$ ），并查卡方分配表的显著临界值。如果计算出的 χ^2

值大于查表的卡方值，则要拒绝虚无假设，而说某个试题在不同的族群上具有 DIF 存在。

上述这种诊断方法，也遭到几种批评：一为即使在某种能力范围内，某两条试题特征曲线没有实质上的差异存在，也会获得很显著差异的试题参数。Linn, Levine, Hastings & Wardrop(1981)便举例说明这种现象也有可能存在，因此容易产生误判的结论。另一为这种卡方分配曲线的统计考验值，只是一种渐近的曲线（也就是说它必须使用大样本才行）而已，它只有在能力参数为已知的情况下，才能适用到试题参数的估计值上，对于要多大的样本才适用？能力与试题参数同时估计的情况下，是否还适用？这种卡方统计值并无法解答这些质疑。

二、比较介于试题特征曲线间的面积

我们曾于前文说过，试题参数不受考生能力分布的影响（亦即具有样本独立的估计特性），因此，根据不同族群考生所估计出来的同一个试题参数或试题特征曲线，在经过衔接或等化之后，这些试题参数应该都已建立在共同的量尺上，其试题特征曲线(ICC)应该会相同，此时，试题特征曲线间的面积应该等于零(Rudner, Getson & Knight, 1980)；如果这些面积不是为零的话，则显示该试题对不同族群考生而言，具有 DIF 的现象。

这种诊断方法的步骤如下：

1. 选取一个适当的试题反应模式。
2. 分别估计不同族群考生的能力及试题参数。
3. 经由衔接的过程，将不同族群考生之能力及试题参数加以衔接，以建立在共同的量尺上。
4. 将能力量尺自 **-10** 到 **+10** 之间，分成 **K** 个等分。
5. 以每个等分的中点为中心，画出该等分的条状长方形图。
6. 计算出每个等分的中点处所能获得的试题特征曲线（机率）值。
7. 计算出两组不同族群考生在每个等分中点处之机率差值的绝对值。
8. 并将该绝对值差值乘上每个等分的宽度（即条状长方形图之宽度），最后，将这些乘积值加总起来。如以数学符号来表示，本步骤可以写成：

$$A_i = \sum_{j=1}^K |P_{1i}(\theta_j) - P_{2i}(\theta_j)| \Delta\theta$$

其中， $\Delta\theta$ 表示每个等分的宽度， $P_{1i}(\theta)$ 和 $P_{2i}(\theta)$ 分别代表两个不同族群考生在某个试题 i 之试题特征曲线（机率）值。

9. 判断 A_i 值，如果 A_i 值很大，则表示试题 i 对不同族群考生而言，具有 DIF 的现象。

上述这种诊断方法也有几项难处：第一，当选用三个参数对数形模式时，如果 C 参数对两组不同族群考生而言不是零或相等的话，则 A_i 值的显著考验便无法进行。第二，由于两组的试题参数都需要估计，因此也需要能力值范围较广的考生加入，所以往往

需要使用大样本；如果每组使用的人数不够多（即能力值范围不够宽广）的话，则容易导致一个错误的 DIF 的结论。

三、比较反应模式与数据间的适合度

如果不同族群考生产生不同的适合度估计值，也表示试题具有 DIF 的现象。这种利用模式与数据间的适合度作为诊断的方法，其步骤如下：

1. 将不同族群考生的数据合并起来，并进行试题与能力参数的估计。
2. 根据估计出的参数值，将每位考生在每个试题上的答对机率值 $\hat{P}_{ij} \quad i = 1, \dots, n; j = 1, \dots, M$ 算出。
3. 计算不同的考生族群在每个试题上的平均 $\bar{\hat{P}}_j$ 值和答对率。
4. 比较各族群在每个试题上的平均 $\bar{\hat{P}}_j$ 值和答对率是否有差别存在，以判定试题具有 DIF 的程度。

上述这种诊断方法也有些缺失，例如，比较不同族群在每个试题上的平均 $\bar{\hat{P}}_j$ 值和答对率的差异时，不论是用卡方考验或 t 考验，都很容易因为使用大样本或大题数而达到显著差异，造成反应模式与数据间的不适合，因而错误下结论说某试题具 DIF 现象。

实例举隅

假设从多数族群（以安格鲁美国人为主）中随机抽取 1000 名受试者当样本，另从少数族群（以土著美国人为主）中随机抽取另外的 1000 名受试者为样本，并从题库中随机抽取 25 个试题给这两个族群样本施测。

假定选用三个参数对数形模式，作为这两族群样本的适合反应模式，并估计出这两族群的试题参数，其中 b 值并予以标准化，以将这两族群的 b 参数建立在同一量尺上。接着，计算出这两族群在每个试题上所夹的面积，以 $\Delta = 0.01$ 为计算单位，算出能力值在 ± 3 之间的面积，并以仿真数据所算出之没有 DIF 情况下之最大分割面积值为 0.498，若每个试题被两个族群的试题特征曲线所夹之面积大于 0.498 时，则该试题被判定具有 DIF，并以*来表示。另外，以 $\chi^2_{3,001}$ 和 $\chi^2_{2,001}$ 作为两种考验试题参数间是否有显著的指标，前者没有把 c 参数列入考虑，后者则有，其分别的临界值为 $\chi^2_{3,001} = 13.82$

和 $\chi^2_{2,001} = 16.27$ 。最后，将这三种诊断结果表列于表一中，其中，标示*者为被诊断出具有 DIF 的试题。

表一 25 个随机试题的试题参数估计值、面积统计数、和卡方值

试题	多数族群	少数族群	DIF 统计数
----	------	------	---------

	b_1	a_1	c_1	b_2	a_2	c_2	面积	$x_{\text{左}}^{1,2}$	$x_{\text{右}}^{1,2}$
1	0.840	0.575	0.190	0.823	0.896	0.170	0.417	5.84	6.01
3	-0.412	0.773	0.190	-0.008	0.906	0.170	0.388	7.90	9.52
5	-1.347	0.413	0.190	-0.953	0.821	0.170	0.609*	21.13*	12.99
8	0.125	0.608	0.190	0.286	0.414	0.170	0.344	5.31	5.21
11	0.319	0.639	0.190	-0.197	0.645	0.170	0.342	17.80*	14.74
13	0.693	0.714	0.190	0.728	0.303	0.170	0.732*	21.86*	19.38*
14	-0.308	1.044	0.190	-0.650	0.551	0.170	0.494	17.12*	15.83
16	-0.193	0.977	0.190	0.286	1.999	0.231	0.405	29.13*	23.07*
20	-0.337	0.536	0.190	-0.106	0.595	0.170	0.238	1.57	2.42
21	-0.514	0.529	0.190	-0.628	0.407	0.170	0.217	2.20	2.22
30	-1.463	0.488	0.190	-0.716	0.839	0.170	0.637*	11.14	9.78
38	-1.168	0.549	0.190	-1.175	0.433	0.170	0.195	4.15	4.64
41	1.011	0.849	0.190	0.943	1.054	0.170	0.214	1.33	1.76
45	1.808	1.166	0.137	2.778	0.509	0.125	0.641*	14.74*	12.08
46	-0.481	0.583	0.190	0.140	0.586	0.170	0.540*	11.62	13.09
49	-0.663	0.661	0.190	-1.128	0.528	0.170	0.290	5.73	3.64
50	0.409	0.431	0.190	0.265	0.430	0.170	0.057	0.56	0.15
52	1.444	1.050	0.190	1.246	1.201	0.137	0.315	1.94	3.19
56	0.338	0.404	0.190	1.545	0.405	0.170	0.880*	14.11*	16.42*
57	0.281	0.685	0.190	-0.497	0.489	0.170	0.536*	32.43*	21.54*
60	0.904	0.569	0.190	1.154	0.531	0.170	0.257	1.19	2.10
64	0.245	0.442	0.190	-0.387	0.280	0.170	0.467	10.52	5.56

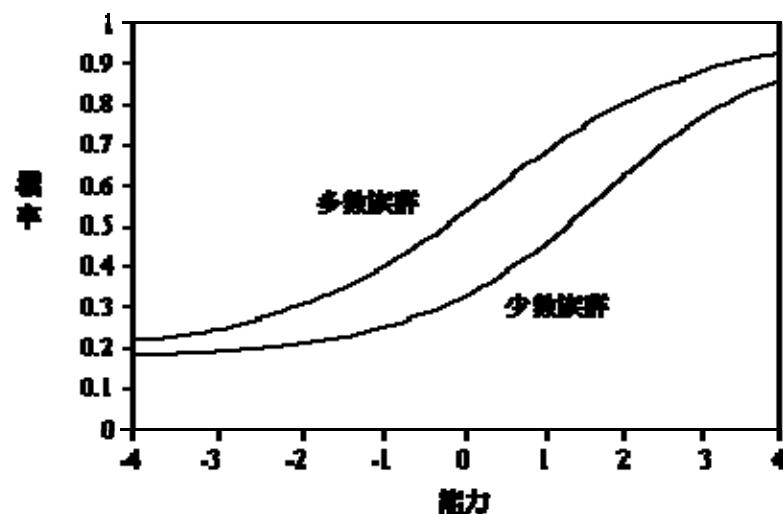
68	-1.398	0.340	0.190	-0.122	0.693	0.170	0.942*	15.41*	15.07
73	-0.567	0.640	0.190	-0.007	1.223	0.170	0.648*	20.29*	20.04*
75	1.646	0.317	0.190	0.534	0.562	0.170	0.722*	23.53*	15.24

a. $\chi^2_{1,001} = 13.82$

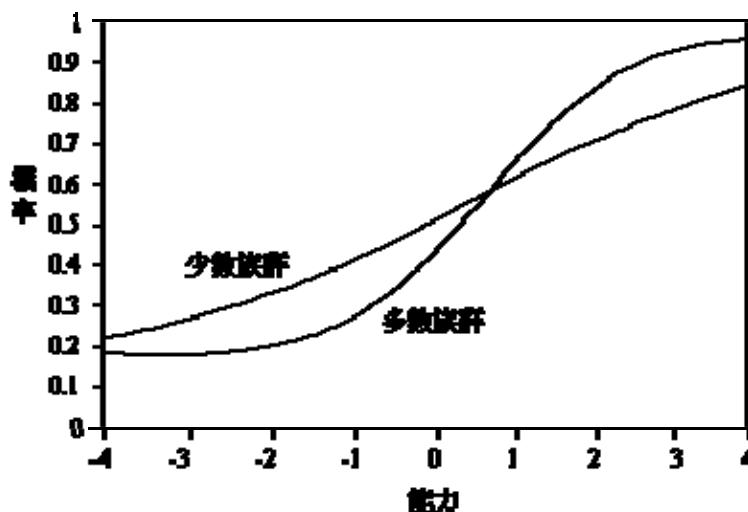
b. $\chi^2_{1,001} = 16.27$

*表示达.001 显著水平

由表一数据可知，前两种诊断方法的一致性达 77%，二者的等级相关系数为 .71。
图一和图二分别是诊断出的 DIF 型态，兹分别说明如下。



图一 多数族群和少数族群在试题 56 上的 ICC 图



图二 多数族群和少数族群在试题 13 上的 ICC 图

由图一所示可知，两个族群在试题 56 上的 ICC 线，多少可以说是平行的，主要的差别只在 b 参数值上，亦即两条 ICC 线的座落位置参数(location parameter)不同而已。这种类型的 DIF 称作「均一变化曲线的 DIF」(uniform DIF)，亦即在所有的能力范围内，这两种族群间的成功机率之差值，是呈均一变化的曲线。

由图二所示可知，两个族群在试题 13 上的 ICC 线表现不同：在低能力量尺的部份，少数族群表现得比多数族群好；而在高能力量尺部份，多数族群却表现得比少数民族还好。这种类型的 DIF 称作「非均一变化曲线的 DIF」(nonuniform DIF)，这时，两个族群在机率上的差异不是呈均一变化的曲线。

由上述表一可知， χ^2_{IRT} 所诊断出的偏差试题数比 χ^2_{ML} 所诊断出者还多，可见后者的诊断方法比前者以及面积统计数法还保守。这种利用 IRT 的程序来诊断试题偏差的一项优点是：这些方法对不同类型的 DIF 极为敏锐。这项特色并非其他非 IRT 程序所能媲美的(Holland & Thayer, 1988; Swaminathan & Rogers, 1990)。但是由上述例子的分析可知，当这些诊断方法所找到的解答不完全一致时，便无法进一步解释其间的差异为什么会有差异存在了(Hambleton, Swaminathan & Rogers, 1991)。

参考书目

1. Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins University Press.
2. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
3. Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum.
4. Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159-173.

5. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
6. Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233.
7. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.

本文转载自研习信息 10 卷（6 期），7-11 页

第十四章

試題反應理論的介紹(十四)

.... 精熟測驗 ...
(Mastery testing)

政大教育系教授 余民宁 着

近二十年来，在心理与教育测验领域里有个重大的改变，那就是效标参照测验 (criterion-referenced tests)逐渐受到重视，并广为流传使用。至今，效标参照测验的用途很广，(1)在军队里，它可被用来评量军人的基本能力；(2)在工业界，它可被用来评定员工工作技能的纯熟度，或评鉴在职训练课程的好坏；(3)在证照考试上，它可被用在各行各业中，以区分出谁是「精熟者」、谁是「非精熟者」；(4)在学校教育中，它可被用来评量学生在某种知识技能上的表现程度。

由于效标参标测验的名辞定义很多（如：Gray (1978)说有 57 种之多），很难予以统一，不过，目前比较被一致接受的定义是：「效标参照测验是指被用来确定个人在某个界定清楚的行为领域中表现程度的测验」(Popham, 1978, p.93)。它又有几个常见的同义词，如：精熟测验 (mastery test)、领域参照测验 (domain-referenced test)、能力测验(competency test)、或基础技能测验(basic skills test)等，本文援用 Lord(1980)的说法，以「精熟测验」一词来显现试题反应理论(IRT)在这方面的应用，并用以区别与古典测验理论所指称的「效标参照测验」的不同，以及介绍精熟测验的整个过程和应用。

精熟测验的设计和编制的步骤

精熟测验的内涵，主要可以分成几个重点：(1)测验试题的设计与选择，(2)测验的计分与报告方式，(3)测验的长度与精熟标准的决定。兹将这整个设计和编制的步骤，条述如下(Hambleton & Zaal, 1991, pp.10-11)：

1. 初步的考虑事项：
 - a. 说明测验的目的。
 - b. 说明该测验所欲测量的目标。
 - c. 说明受试者的特性及特殊的施测设备。
 - d. 初步决定试题格式（如：客观测验的试题或实作导向的试题）。

- e. 决定编制测验所需的时间和成本。
 - f. 慎选合格且适当的命题委员（如：考虑他个人的专长或在发展测验中所扮演的角色的重要性）。
 - g. 说明初步的测验长度（如：要测量那些目标、需要多少题数、及施测时间多长）。
2. 审视测量的目标
- a. 审视测量目标的陈述是否明确清楚，目标的适当性可否被接受。
 - b. 选择测验所欲测量到的一组目标。
 - c. 针对每项目标描述所需试题的特征，并且审视这些特征的完整性、正确性、明确性、和实用性。
3. 撰写试题
- a. 撰写大量试题，以作为预试之用。
 - b. 输入计算机化题库，以便利修改和存取（参见本系列论文之（十一）——题库的建立）
 - c. 进行试题初步编辑工作。
4. 评量内容效度
- a. 延聘一批课程、学科、与测量专家。
 - b. 请这批专家评阅这些试题是否符合它们所欲测量的目标、是否具有教材内容的代表性、以及是否不受刻板印象的影响。
 - c. 审视这些试题，以判定其技术上的适切性。
5. 修改试题
- a. 有必要时，根据上述 4b 到 4c 的步骤，修改试题或删除不适当的试题。
 - b. 如有需要，撰写补充的试题，并重复上述第 4 个步骤。
6. 初步预试
- a. 编辑试题成试卷的形式，以便进行预试。
 - b. 针对一群适当的考生施测。
 - c. 进行试题校准和试题偏差的诊断（参见本系列论文之（十三）——试题偏差的诊断）。
7. 测验试题再修改
- a. 根据 6c 的步骤，如有必要，需对试题再加以修改或删除。
8. 组合成正式的测验
- a. 决定测验的长度、所需的题型数目、及每个目标需多少试题数。
 - b. 从上述有效的候选试题中（多半是由题库中），挑选所需要的适当试题。
 - c. 准备测验指导语、练习用的试题、测验题本、计分卡、答案纸……等。
 - d. 补充说明指导语不清楚的地方、考生作答有那些注意事项、特殊考生（如：残障考生）所需的作答时间等。
9. 设定精熟的标准
- a. 决定考生表现程度的描述或精熟程度的设定，是否能够符合测验的目的（如果该描述是主要的用途的话，则跳到第 10 个步骤）。
 - b. 说明设定区分为「精熟」与「非精熟」之标准的挑选过程；如果必要的话，设定一个以上的标准（如：分成「卓越」、「优良」、「尚可」等）。
 - c. 说明残障考生所适用的特殊标准。

d. 说明需要重测的考生的另一种计分方式。

10. 正式预试

- a. 设计施测的方式，以便收集测验分数的信度与效度等方面的信息。
- b. 对挑选出的一群适当的考生进行施测。
- c. 评估那些为符合特殊需求而改变之施测过程所可能造成对测验的信度和效度估计的影响。
- d. 评量施测程序、测验试题、及测验分数的信度和效度。
- e. 根据上述所获得的技术性资料，进行最后的修改。

11. 准备使用手册

- a. 准备一份施测者或监考者须知手册。
- b. 准备一份技术性使用手册。

12. 收集额外的技术性资料

- a. 进行信度和效度的分析研究。

测验的长度

有关精熟测验应该具备多少试题才算适当的研究，一直是个很数量化的研究课题，所累积的研究文献也很多(Hambleton, 1984)。本文仅列举较具实用性的一种如下。

当我们确定测验分数的用途，也对其用法加以描述后，有关考生得分的专精分数估计值（又称作「领域分数」(domain score)）的测量精确度，可用公式表示如下：

$$(\text{精确度})^2 = \frac{\hat{x}(1-\hat{x})}{n} \quad (\text{公式一})$$

其中，专精分数估计值 \hat{x} 可以表示如下：

$$\hat{x} = \frac{1}{n} \sum_{i=1}^n P_i(\theta) \quad (\text{公式二})$$

$P_i(\theta)$ 为具有能力估计值为 θ 的考生在试题 i 上答对的机率；所有测验试题答对机率之和，即为该考生的真实分数(true score)。因此，所谓的专精分数即是真实分数的平均数；亦即是考生答对某种目标领域的测验内所有试题的机率。该值为一比率分数，其值域介于 0 与 1 之间；其值愈接近 1，表示该考生的精熟程度愈大，反之，该值愈接近 0，则表示该考生在测验上的表现反应愈不精熟。

由公式一可以推论出适当的测验题数应该是：

$$n = \frac{\hat{x}(1-\hat{x})}{(\text{精确度})^2} \quad (\text{公式三})$$

例如，假设已知某群考生的专精分数为 .80，且我们希望该专精分数估计值的精确

度至少能够达.10 的话，则将此二数值代入公式三，可以获得：

$$n = \frac{.80(1-.80)}{(.10)^2} = 16$$

换句话说，我们若想要使某群考生在某个测验上的专精分数达到.80，且其估计值的精确度也达.10 的水平的话，则我们必需要编制出一份至少含有 16 个试题的测验，才能符合我们所需要的测验目的的要求。由此可见，要编制出一份达到某种测验目的的精熟测验，其题数的多寡完全取决于专精分数和估计精确度两个因素：专精分数愈接近.50，所需之题数则愈多；专精分数愈接近于两极端（即 0 或 1），则所需题数愈少。若所要求的精确度值愈大，则所需的题数愈少；若所要求的精确度值愈小，则所需的题数便需要愈多。

如果精熟测验分数的目的是用来区分「精熟者」与「非精熟者」的话，则可用来帮助决定题数多寡的参考依据就更多。Millman(1973)和 Wilcox(1976)提供了许多对照表，可用来帮助决定适当的测验长度、专精分数、通过分数(passing score)或标准设置(standard setting)等问题。

精熟标准的决定

在精熟测验中，有关通过分数等标准设置问题之研究文献，可说是已经到了汗牛充栋的地步。

根据数字学者(Berk, 1984, 1986; Hambleton, 1990; Hambleton & Zaal, 1991)的归纳，有关标准设置之研究方法，大致可以归纳成三大类，大类内各有数种较有名的方法：

一、判断的方法(judgmental methods)

这种方法主要是聘请专家评审每一个试题，以判别出最低能力考生所可能表现到什么样的最佳程度。这类设定通过标准的方法，有三种较为常用的方法较有名，分别是：

1. Nedelsky 法：首先，请个别的专家找出最低能力考生在选择题的诱答选项中，能够删除（或以消去法消除）的选项数目。因此，该试题的最低通过标准即订定为剩余未被删除之诱答选项的数目之倒数，即为最低能力考生在该试题上的「猜测分数」(chance score)。再将每个试题之最低通过标准（即猜测分数）加总起来，便得此一测验的通过标准。若有数字专家进行判断，则以个别之通过标准之和的平均数，作为该测验之通过标准。
2. Ebel 法：根据试题的相关性和难度两个向度，请专家进行评定。其中，相关性分成四个水平，难度分成三个水平，共形成 4×3 的列联表，再请专家就每一细格中，最低能力考生所可能答对之百分比进行评定。再将数字专家评定一致之细格题数加总除以总题数，便得此一测验之通过标准。
3. Angoff 法：请专家就每一个试题中，最低能力考生所可能答对之机率，进行评定。将每题可能答对之机率加总，便成为该专家所判断的通过标准。再将数字专家之判断的通过标准加以平均，便成为该测验之最后的通过标准。

二、实征的方法(empirical methods)

这种方法是以考生实际的作答数据的分析结果，作为设定通过标准之依据。又可分成：

1. Livingston 法：从外在选择一个效标(criterion)，并建立一条直线的效用函数(linear utility function)，以决策理论的方法找出能够使该效用函数达到极大的分数切割点(cutoff score)，便是该测验的通过标准。
2. Linden & Mellenbergh 法：找出能够使「期望的损失」(expected losses)（即分类错误的代价）达到最小的分数切割点，便是该测验所需之通过标准。若此点找出后，将使效标分数大于此点以上者（即被判定为精熟者）能够通过测验；效标分数小于此点者（即非精熟者）无法通过该测验。

三、混合的方法(combination methods)

这种方法乃揉和上述两个方法，用来设定通过标准的一种过程。又分为：

1. 边缘组法(borderline-group method)：首先要求专家对每一教材内容的最低可接受的表现程度作一定义，再列举一批表现水平接近此一划分为精熟与非精熟的边缘线的考生，然后编辑测验对此批考生施测，取其得分之中位数(median)，作为该测验之通过标准。
2. 对照组法(contrasting-group method)：首先要求一批专家定义精熟某教材内容的最低可接受的表现程度，再找出他们已确知某些已达精熟和未精熟的考生。针对此二组考生施测，并将此二组考生的得分分配曲线，一一画在每个目标范畴图上，取其两线的交叉点作为起始的通过标准(initial standard)。然后，再渐次调整该交叉点，使分类错误率达最小的位置为止，此时的决定点即为最后的通过标准。

总之，标准的设定终究还是属于判断的历程，需要参与者(1)熟悉教材内容和各种设定标准的方法，(2)有评定试题表现和测验分数分配曲线的能力和经验，(3)以及了解使用该测验的社会与政治背景。如此才能有个良好、公正的标准诞生(Hambleton & Powell, 1983)。

精熟测验的未来发展方向

精熟测验的编制技术、应用、与改进，均已日臻成熟的地步。目前，它仍在研发的领域有：(1)标准设置的方法，(2)改进测验分数使用效果的分数报告格式，(3)以及描述目标的方法。未来，尚可改进精熟测验的实用性和提高未来的运用潜力的方向，计有(1)配合微电脑的使用，研究如何存取、施测、和计分，(2)配合试题反应模式，研究发展目标、测验试题、和考生可以随时参考使用的各种继续成长或发展的量表。有关精熟测验的编制、应用、和发展方面的教科书和使用手册，读者可参阅Berk(1984)、Hambleton(1990)、和 Popham(1978)等人的专著。

参考书目

1. Berk, R. A. (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

2. Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
3. Gray, W. M. (1978). A comparison of Piagetian theory and criterion-referenced measurement. *Review of Educational Research*, 48, 223-249.
4. Hambleton, R. K. (1984). Determining test lengths. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction*. (pp. 144-168). Baltimore, MD: The Johns Hopkins University Press.
5. Hambleton, R. K. (1990). *A practical guide to criterion-referenced testing*. Boston, MA: Kluwer.
6. Hambleton, R. K., & Zaal, J. N. (Eds.) (1991). *Advances in educational and psychological testing*. Boston, MA: Kluwer.
7. Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation and the Health Professions*, 6, 3-24.
8. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
9. Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 43, 205-216.
10. Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.
11. Wilcox, R. (1976). A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1, 359-364.

本文转载自研习信息 11 卷（1 期），7-11 页

第十五章

試題反應理論的介紹(十五) ---- IRT 的其他應用 --- (The application of IRT on other fields)

政大教育系教授 余民宁 着

本系列介绍专文的前七篇（即之一到之七）是讨论试题反应理论的理论部份；而接下来的七篇（即之八到之十四）是讨论试题反应理论的应用部份，包括：测验的编制、等化、题库、计算机化适性测验、偏差试题的诊断、及精熟测验等，这些应用范畴都是心理计量学领域中原本就有的。除此之外，试题反应理论也被应用到其他领域，与其他学科知识相结合，作为研究与改进测量的工具之一。底下所介绍的应用课题，

便是它所被重视的地方。

量表的翻译与修订

教育与心理测验经过编制使用后，过一段时间再检讨它们，研究者也许会发现已有许多试题的测量特征(measurement characteristics)发生改变，变得不再适合目前使用上的需要；在这种情况下，研究者通常会针对这些不合时宜的测验或量表重新加以修订(revision)。

有些时候，研究者为了研究上的需要，必须编制一份新的测验。他除了可以依照测验编制过程来进行编制外，多半他会参考原文版本的测验或量表，将它们翻译成中文版本，再进行修订，以节省重新编制的时间。在此，不论是经过翻译再修订，或直接修订不合宜的测验或量表，都必须保留或维持新修订的试题具有所希望的测量特征，达成用户所期望的测验目的。因此，试题反应理论的技术可以派上用场，使得新修订的试题都能具有研究者所希望的测量特征。

翻译外文量表或测验所可能遭遇到的最大难题，便是泛文化差异(cross-cultural difference)所产生的泛语言间的语意不同问题，尤其是心理学上的语意问题。因此，研究泛文化问题的学者在需要翻译量表时，多半会采用「倒译法」(back translation)来进行量表的翻译：亦即将原文译成中文，再请专家根据中文译成原文，比对这译文与原始的原文在语言学上的正确性与差异性，修正有差异之处，再持续这反复译文的比对工作，直到语意皆正确无误为止(Brislin, 1970, 1980)。

为了验证每道试题在不同文化背景下的受试者群体间是否有所偏差(bias)，古典测验理论学者(Gulliksen, 1987; Lord & Novick, 1968)多半会比较每道试题在每种群体下所获得的一些指标：(1)难度值(即每道试题的正确反应比率)；(2)鉴别度值(即试题得分与测验总分间的点二系列相关系数)；(3)考生得分的平均数；(4)考生得分的标准偏差；(5)试题反应共变量的因素分析等(Hulin, Drasgow & Parsons, 1983)。然而，前四项指针都属于是样本依赖(sample dependent)的指针，会随着样本母群的不同而不同，因此，并不适用于泛文化间的比较；而后者指标，往往需要有其他基本假设(如：符合常态性等)的先前条件，因此也不太适合二元化分类的试题反应资料。所以，为了克服上述这些缺点，也唯有仰赖试题反应理论的分析技术了。

试题反应理论在修订一份翻译量表上的应用，包括试题参数的估计具有不变性、讯息函数、量尺的等化、以及偏差试题的诊断等(参见本系列论文之四、之七、之九、之十、及之十三)特性及方法的使用，可说是一种综合性的应用。归纳起来，试题反应理论在这方面问题的应用程序，包括下列诸项(Drasgow & Kanfer, 1985; Hulin, Drasgow & Parsons, 1983; Hulin, Drasgow & Komocar, 1982; Hulin & Mayer, 1986)：

1. 挑选懂得双语(bilinguals)的受试者和仅懂单语(monolinguals)的受试者作为施测的对象，分别给予原文版和中文版译文的测验。
2. 针对这批测验数据进行校准(calibrations)的工作，其中以双语受试者为定锚受试者，因为他们共同接受这两种版本的测验。
3. 针对每道试题，进行其试题特征曲线(即 ICC 线)的比对，或计算其所夹面积之大小，以诊断出偏差试题。
4. 针对有偏差的试题，再进行倒译法的语意修正，期使语意能适用于不同的文化。修正后的试题再予以单语受试者施测，再进行校准，再比对 ICC，直到没有偏

差试题出现为止。

5. 最后，根据原文版的试题特征为基础，将中文版译文试题特征等化成相同的量尺。

如此一来，新译版的测验或量表与原始量表皆具有相同特征的量尺单位，故可以视为原始量表在国内使用。如果为了严谨起见，也可以仿同古典测验理论在编制测验时作法，进行因素分析、试题分析、信度与效度分析等过程，期求所修订的测验或量表更臻完备。

由此可见，修订一份翻译量表的最终目标在于建立「等值的量尺」(equivalent scale)，即将新译量表建立在原始量表的量尺上，以避免译题受文化偏差的影响，并同时能享有与原版试题同样的测量特征。有关泛文化的测验问题，有兴趣的读者可以参阅 van de Vijver & Poortinga(1991)的论文。

适度性测量

在大规模的施测情境中，尤其是使用单选题(multiple choice item)作为测验试题时，研究者往往你会发现考生得分有下列各种奇怪现象产生，例如：

1. 在施测前，考生死背某些很困难试题的答案，故能顺利答题，远超乎他能力所及；
2. 高能力但低语文程度的考生，他的答题结果反映不出他的真正能力；
3. 特别有创意的高能力考生发现有新颖的解题方法和解释，但却被当成错误来计分；
4. 考生漏掉部份试题，致使后续的答案都填在不正确的位置上，导致一连串错误；
5. 某些低能力的考生过度猜对远非他能力所及的试题数；
6. 某些考生的应试技巧过度保守，若非十分把握的试题，绝对是空白不答；
7. 某些考生过度不熟悉测验格式，导致胡乱猜题，甚至胡乱作答或以作弊方式来应付施测。

综合上述现象，研究者因此会怀疑(1)考生的得分可能不是代表能力的一个适当测量值，(2)由试题与试题所构成的答题组型，可能会呈现不寻常(unusual)的状态。因此，上述这些原因都会产生不寻常的答题组型(或称反应组型(response pattern))，使得考生的得分不再代表他真正的实力。为了能够从考生的反应组型中，找出这种代表不适当得分的不寻常反应组型来，于是便有「适度性测量」(appropriateness measurement)的研究诞生(Levine & Rubin, 1979; Levine & Drasgow, 1982)；简单的说，适度性测量便是用来找出这种不适当测验得分(inappropriate test scores)的方法。

Levine & Rubin (1979, p. 271)对适度性测量作了一个定义：「适度性指标(appropriateness index)是一种对某个心理计量学模式与考生的反应组型间之适合度(goodness of fit)的简单测量。如果考生在答案卷答题结果与其能力相仿的话，适度性指标应该会较高；反之，若答案卷上答题结果愈不像考生应有的能力作为时，则适度性指标应该会较低。就像考生的测验得分一样，考生的适当分数应该只是他在试题上的得分的唯一函数而已。因此，适度性指标是考生在答案卷上答题情形的内在证据的显示指针，反映出他的答题结果是否与其他具有相同能力的人的答题结果相一致。」所以，适度性指标便是用来测量考生的反应组型是多么不寻常的程度，典型的考生反应组型若偏离所期望的反应组型，则适度性指针便能显示出他的不适当测验得

分，而不是用来矫正造成这种不适当分数的原因。

适度性指标有三种(Levine & Rubin, 1979)，各是以一种数量指标来表示，分别是：

1. 边缘机率(marginal probabilities)指标：此指标适用于一般正常的考生行为；即以某特定能力群体（如常态分配）中随机抽取之考生的反应组型的条件机率为基准，求出同属这一能力分配的群体考生的平均数，用以代表该考生之特定的反应组型的边缘机率。当某位考生的反应组型出现不寻常时，即异常组型(aberrant pattern)（如：高能力考生答错简单的试题，或低能力考生答对困难的试题所形成的反应组型），他的边缘机率值便可能相当的低；反之，则否。
2. 近似比值(likelihood ratios)：是以标准的近似比值分析技术，来针对通用的试题反应模式与另外的类化模式在适合用来分析考生的反应组型数据中，允许每个答题所需之能力都可以有所不同的情况下，各分别求出这两种模式在该考生之反应组型上的机率的最大值，并比较其间的差值，看看该类化模式(generalized model)的适合度是否比通用的模式更好。
3. 估计的能力变异量值(estimated ability variation)：在允许每个答题所需之能力都可以有所不同的情况下，分别估计出模式的能力参数值及其估计值的变异数，并以此估计值作为衡量异常程度(degree of aberrance)的指标，故又称为异常程度估计值。

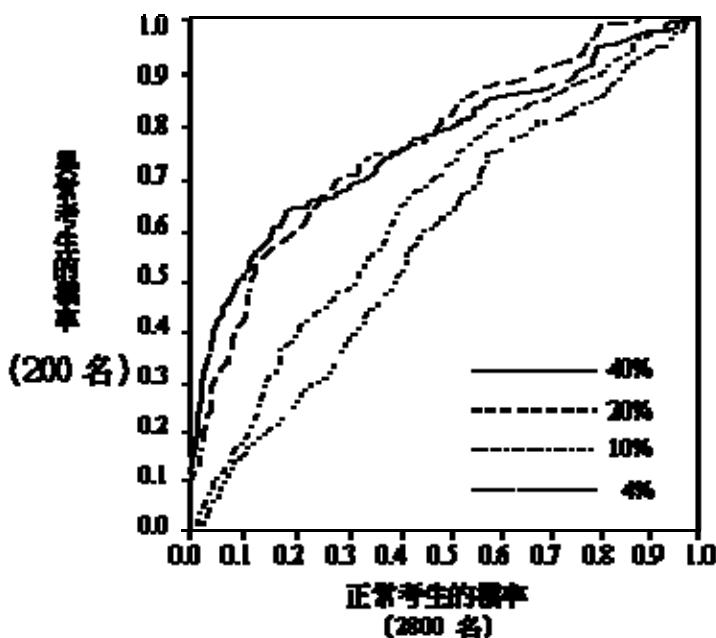
适度性测量便是利用上述指标来作为表示异常反应组型的指标，它包含两个过程：(1)试题参数的估计，或称为测验规准(test norming)；(2)指标的计算，或称为个人测量(person measurement)。这些过程很类似试题反应理论中对测验数据所进行的校准工作，不过是以上述三种指标作为找出异常反应组型的指标。在评定和诊断一个测验得分是否为不适当，研究者往往必须先决定一个效标值，以作为判断的参考值，并利用统计决策理论中的接受者操作特征(receiver operating characteristic)（简写成 ROC）曲线，来提供决定此一分割点的依据。当研究者以适度性指标来进行考生得分是否为适度的归类时，他事先决定一个切割的效标值（假设以 t 来表示），然后将异常考生正确归类为异常的百分比，与将正常考生错误归类为异常的百分比，分别表示如下：

$x(t)$ ：适度性指标值小于 t 的正常考生的百分比。

$y(t)$ ：适度性指标值大于 t 的异常考生的百分比。

而所谓的 ROC 曲线，便是根据各种可能不相同的效标 t 值，将每对 $(x(t), y(t))$ 百分比值相对应画成的曲线分布图。其中， $x(t)$ 值称错误的警示比率(false alarm rate)， $y(t)$ 值称为正确分类比率(hit rate)。在大多数的应用情境中，太高的错误的警示比率是不被允许的，因此，我们需要适当的 t 值，使得正确分类比率值 $y(t)$ 较大，而错

误的警示比率值 $x(t)$ 较小(Levine & Drasgow, 1982)。



图一 以每组 200 名考生的异常程度指标（边缘机率值）所画成的 ROC 曲线

图一所示为根据 3000 名模拟的考生反应组型资料所画出的 ROC 曲线，其中横轴 (即 $x(t)$) 代表适度性指标小于效标 t 值的正常考生的百分比，而纵轴 (即 $y(t)$) 则代表适度性指标小于效标 t 值的异常考生的百分比，而 ROC 曲线即为 $(x(t), y(t))$ 点组合所构成的分布曲线。此 ROC 曲线有个简单的判别方法：即一个不良的适度性指针会使得 ROC 曲线愈接近 $x = y$ 的对角线，而一个较佳的适度性指针会使得 ROC 曲线偏离在对角线之上。图一所示即为各种假想的低能力群考生，在各种异常组型百分比的假设下，所被画出的 ROC 曲线；由此曲线图可知：异常组型的百分比愈大者，愈容易显示出来，如图中 20% 者就比 4% 者明显偏离对角线，显示前者愈容易被诊断出来。

适度性指标被用作诊断考生的异常反应组型（显示在其不适当的测验分数上），已被证实获致良好的成效。有关这方面的理论与应用的报告，有兴趣的读者可再参阅 Drasgow(1982)、Drasgow & Guertler(1987)、Drasgow, Levine & Williams(1985)、Drasgow & Levine(1986)、及 Levine & Drasgow(1988)等论文。

参考书目

1. Brislin, R. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1, 185-216.
2. Brislin, R. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Methodology* (Vol. 2, pp. 389-444).

- Boston, MA: Allyn & Bacon.
3. Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6, 297-308.
 4. Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, 72, 10-18.
 5. Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70, 662-680.
 6. Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
 7. Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, 10, 59-67.
 8. Gulliksen, H. (1987). Theory of mental tests. New York: Wiley.
 9. Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL: Irwin.
 10. Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, 67, 818-825.
 11. Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, 71, 83-94.
 12. Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies,. *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.
 13. Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161-176.
 14. Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
 15. Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
 16. Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 277-308). Boston: Kluwer Academic.

第十六章

試題反應理論的介紹(十六)

.... IRT 的未來 ...
(The Future of IRT)

政大教育系教授 余民寧 着

测验理论的发展有两个趋势——一为「理论的发展愈趋向数学化」，另一为「理论的应用愈趋向计算机化」（余民宁，民 82，民 82）；这意谓着，将来的测验理论用户必须兼备数学与计算机的良好训练，才能对当代的试题反应理论的了解与应用驾轻就熟。

自从 Lord(1980)发表第一本以「试题反应理论」为名的专书后，当代的测验理论才正式正名成立，以后的发展至今，可说是如火如荼。表一所示，是有关试题反应理论方面的专书或论文集专著。由表一可见，试题反应理论中的重要概念，早在 Lord(1980)发表专书之前就已存在，不过，近年来随着计算机科技的发展，试题反应理论的演进，有愈来愈快速的趋势。

表一 有关试题反应理论发展的专书或论文辑

作者 (年代)	书名
Rasch (1960 / 1980)	Probabilistic models for some intelligence and achievement tests.
Lord & Novick(1968)	Statistical theories of mental test scores.
Wright & Stone(1979)	Best test design.
Jensen(1980)	Bias in mental testing.
Lord(1980)	Applications of item response theory to practical testing problems.
Weiss (Ed.) (1980)	Proceedings of the 1979 computerized adaptive testing conference.
Andersen(1980)	Discrete statistical models with social science applications.
Wright & Masters(1982)	Rating scale analysis.
Berk (Ed.) (1982)	Handbook of methods for detecting test bias.
Wainer & Messick (Eds.) (1983)	Principals of modern psychological measurement.
Weiss (Ed.) (1983)	New horizons in testing: Latent trait

	test theory and computerized adaptive testing.
Hulin, Drasgow, & Parsons(1983)	Item response theory: Application to psychological measurement.
Hambleton (Ed.) (1983)	Applications of item response theory.
Berk(1984)	A guide to criterion-referenced test construction.
Embreton (Ed.) (1985)	Test design: Developments in psychology and psychometrics.
Baker(1985)	The basics of item response theory.
Hambleton & Swaminathan(1985)	Item response theory: Principles and applications.
Crocker & Algina(1986)	Introduction to classical and modern test theory.
Wainer & Braun (Eds.) (1988)	Test validity.
Langeheine & Rost (Eds.) (1988)	Latent trait and latent class models.
Linn (Ed.) (1989)	Educational measurement (3 rd ed.)
Freedle (Ed.) (1990)	Artificial intelligence and the future of testing.
Suen (1990)	Principles of test theories.
Wainer et al. (1990)	Computerized adaptive testing: A primer.
Hambleton(1990)	A practical guide to criterion-referenced testing.
Hambleton, Swaminathan, & Rogers (1991)	Fundamentals of item response theory.
Hambleton & Zaal (Eds.) (1991)	Advances in educational and psychological testing.
Baker(1992)	Item response theory: Parameter estimation techniques.
Frederiksen, Mislevy, & Bejar (Eds.) (1993)	Test theory for a new generation of tests.
Holland & Wainer(1993)	Differential item functioning.

近年来，试题反应理论正朝下列几个方向发展，其中亦隐含许多未来发展的契机：

(一)计算机化适性测验

计算机化适性测验 (computerized adaptive testing, 简称为 CAT) 的发展与

测试，至今可说是相当完备与周全。举凡纸笔测验(paper-and-pencil tests)所具有的功能、特性、优点、或测验的结果，CAT 均具有，并且，CAT 比传统的纸笔测验还节省一半左右的施测时间。这就是 CAT 迷人之处，也是它受到重视的原因。

然而，这些特性并不表示 CAT 到此已无改进的余地。下列几项实际的课题，正是 CAT 未来所要改善的重点及发展方向(Wainer et al, 1990):

1. 时限(time constraint)问题：原本 CAT 的一项优点是：让受试者根据自己的反应速度作答，因此，在施测时限上非常具有弹性。然而，有些受试者若犹豫过久、或不安于施测情境，则他们的作答时间将会拖得很长一段时间，这对实施 CAT 而言，也就无法发挥 CAT 的特性了。所以，针对每个试题或每份测验的作答时间予以适当的限制，还是有必要的。
但是，设立时限却会引发另外两种问题：一为每个试题或每份测验的作答时间还剩多少，必须显示让受试者知道；另一为如何去计算那些尚未答完的测验得分。这两个实际问题很难解决；也许未来的 CAT 系统宜朝增加个计时功能的装置，或设置一些扣分的计分方法，方能克服 CAT 的发展瓶颈。
2. 作弊及其他不适当的考试行为：就如前文中所述，受试者在施测情境中，可能会有一些不寻常的反应组型(unusual response patterns)出现，这些不寻常的作答行为可能是来自：作弊(cheating)、粗心大意、焦虑过度、或任意猜题等因素。不过，自从 Levine & Rubin (1979) 发展出适度性测量(appropriateness measurement)后，受试者的不寻常反应组型已能被适度性指标所诊断出。未来的 CAT 走向，似可将适度性测量的诊断系统包含在 CAT 系统里，以提供诊断与补救措施，使 CAT 更能发挥适性测验的功能。
3. 省略(omitting)问题：在 CAT 中，不太可能会有省略的试题出现，因为每位受试者必须在回答一个试题后，CAT 才会挑选下一个试题呈现给受试者作答，因此，受试者不会错过任一试题。但是，有些受试者在纸笔测验中，面对自己没有把握的试题时，通常会采取略过的作法，等其他试题均回答完毕后，再回头尝试作答刚才省略的试题。这种作法在 CAT 中却无法办到，因为受试者必须在屏幕上回答一个试题，才能有机会回答下一个试题；换句话说，若受试者对目前屏幕上所呈现的试题没有把握，而想暂时跳过，等待后来再回头作答时，CAT 却无法允许受试者有此选择，受试者还是得强迫作答后，才能回答下一个试题。因此，未来的 CAT 系统宜朝多增加一个选项：「跳至下一个试题」来设计，但是，这又引发一个问题：「该多出来的选项应该如何计分？」如果把省略的试题当成是答错，则受试者很可能逃避他没有把握者，而选答他完全会的，因此，最后的计分结果也很可能不正确或不公平。也许 CAT 系统亦宜将分数等化方法包含在设计里，以期获得公正、公平的计分。
4. 在人格与态度测量上的应用：目前，IRT 在人格与态度测量(personality and attitude measurement)上的应用，远不如在教育测量上的应用。近期的社会或人格方面的研究文献显示，仅有少数几个应用 IRT 的方法，来解决调查问卷方面的资料分析问题（如：Thissen & Steinberg, 1988 等）。因此，发展 CAT 系统并应用到这些研究领域，是颇值得开发与尝试的新研究课题。不过，在从事人格与态度测量方面之研究时，所常遇到的问题：「原本设计用来测量某些人格与态度变项的试题，会随着时间的流逝，而在语意、与测量结构的关系、或极端选项的水平上，产生很明显的变化」，也会发生在应用 IRT 方法的 CAT 系统中，此时的问题便是：「试题参数漂流」(item parameter

drift) (Bock, Muraki, & Pfeiffenberger, 1988) 的问题，亦即是试题参数估计值随着时间而改变的现象。因此，未来的 CAT 在这方面的研究，也需将 IRT 所用的量尺随时更新及校准，以增进测量的精确度。

(二) 认知诊断测验

在过去十余年来，认知心理学(cognitive psychology)的发展已逐渐蔚为心理学的主流，它的研究方法也已摒弃过去主观式的投射分析，而逐渐改采较客观、可以量化、和较深奥的数学模式为基础的研究架构，来探究人类的学习行为及愈来愈复杂的认知行为。其中，对教育界影响较多的便是针对人类学习中「认知失误」(cognitive bugs)行为的研究，尤其在与人工智能(artificial intelligence)的结合，逐渐兴起一股诊断测验(diagnostic testing)学的新兴研究领域；当然，这门新的研究领域是以试题反应理论为基础，才足以彰显它的重要性和未来的潜力。

Tatsuoka (1983, 1986, 1990) 及 Tatsuoka & Tatsuoka (1987, 1988) 发展出一种叫做「规则空间」(rule space)的数学模式，用来诊断及侦测小学生在解决算术中四则运算之问题时，为何有的学生会答对？而有的学生会答错？而且答错者的反应组型皆不相同之原因。他们发现学童使用错误规则(erroneous rules)来解题，因此产生系统化的错误，这些错误反映在学童的不寻常反应组型里，同时，他们导出这些观察得到的失误情形的理论分配，并且称这个分配为「失误分配」(bug distribution)。这项重要的发现，对教育的实务问题具有很重大的涵意：教师可以透过有目的、结构化的设计试题，经过规则空间的分析，就能顺利找出或诊断出具有认知失误、或「错误概念」(misconception)的学生来，以便对症下药进行补救教学。

然而，Tatsuoka 的研究并非没有限制，Linn(1990)就曾评论说：「Tatsuoka 的研究仅限于结构良好(well-structured)的问题领域，至于结构较不良的问题，规则空间是否仍能提供有用的讯息，则有待日后的证实」(P. 491)。因此，发展一套分析技术，以期能适用于各种学科领域知识的诊断，是未来认知诊断测验可以走的方向。

近年来，认知心理学的研究结果给行为科学家更多的启示，人类的学习行为也以计算机仿真方式呈现给人们了解，致使许多领域的学者结合起来一起研究，因此，认知诊断测验也逐渐结合认知科学、教学研究、及心理计量学而成为一门新科学；甚至认为诊断测验与教学是一体的，不可单独分开处理(Embretson, 1990; Marshall, 1990)。也因为如此，有些心理计量学者(Mislevy, 1993; Lohman & Ippel, 1993; Snow & Lohman, 1993)甚至开始主张「新的测验理论」诞生——以认知理论为基础的新的评量方式和测验设计方法。由此可见，未来的认知诊断测验的新走向也许是：根据某种认知科学的理论为基础，依据该理论设计新型的诊断测验试题，再提出可能评量该理论模式的 IRT 测量模式，以验证该理论下的评量是否成立，并予以认知、测量、或教育领域中有意义的结果解释。

(三) 多向度的 IRT 模式

目前所盛行的 IRT 模式，都是属于单向度的模式。然而，在实际的心理与教育测量上问题，却很少是单向度，而多半是多向度的(multidimensional)。因此，朝多向度 IRT 模式发展（不论是二元计分法的或多元计分法的），可能是未来 IRT 应该走的路。

多向度 IRT 模式首由 Lord & Novick(1968) 和 Samejima(1974) 提倡其理论概念，后又有两位学者(Embretson, 1984; McDonald, 1989) 尝试不同形式的模式。多向度模式能够提供更佳适合目前测验数据，并且提供试题与受试者能力有多个向度

表征的新机会。至于多向度模式的参数值能否适当的估计出，或试题与受试者能力的多向度表征是否对实务上有所用途或帮助，这可能需要等待进一步研究方能得知。由此可见，研发多向度 IRT 模式、设计其适用的计算机软件程序、以及朝实务界的应用，可能是未来 IRT 该走的新方向。

(四)潜在类别模式

IRT 所发展出来的模式，多半是假设受试者的能力参数的特性是连续的。然而，有些实际资料的适合度问题指出，在某些情况下，这种假设也许不是很恰当，因此才造成模式与数据间的不适合问题。如果我们将该假设予以放宽，允许能力参数的特性不是连续的，而是间断的，也许可以解决并解释模式与数据间不适合的现象问题，这也就是另一种新的 IRT 模式——潜在类别模式(latent class model)的研究起源(Rost, 1985, 1988)。

潜在类别模式与试题反应理论(或又称为潜在特质理论)所用的模式间最大差别，在于彼此对考生之能力参数的属性的假设不同：前者假设为间断的，后者假设为连续的。这两者所使用的数学模式，都是一样的深奥、复杂，都是以机率的观念来表示某位具有某种能力(或能力类别)的考生在某个试题上答对之可能性。

潜在类别模式所能应用到实际的测验数据上的情况还不多，不过，未来的许多研究仍可使用潜在类别模式，尤其是在人格与态度测量上。例如，其他能够适用于评定量表(如：Likert 氏的五点量表)数据的所有测量数据的分析，除了可用传统的分析方法外，潜在类别模式的分析也许是未来的新秀。

自从 Lord(1980)提出试题反应理论一词以来，试题反应理论的发展至今仍方兴未艾。未来 IRT 发展新走向，除了可以朝上述四个方向及其改进建议迈进外，亦可朝应用面来进行：建立全国性大规模的教育测量及评量标准，并且将应用成果落实在实际教育问题的解决上。这些未来的发展方向，是全体心理测验学者及教育人士所需继续努力的。

(本系列介绍论文完)

参考书目

1. 余民宁(民 82)。测验理论的发展趋势。政大心理系承办「心理测验之学术及实务研讨会」之论文宣读。
2. 余民宁(民 82)。测验理论的发展趋势。载于中国测验学会主编：心理测验的发展与应用(23-62 页)。台北：心理。
3. Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275-285.
4. Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175-186.
5. Embretson, S. E. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N Frederiksen et al. (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 407-432). Hillsdale, NJ: LEA.
6. Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4,

- 269-290.
7. Linn, R. L. (1990). Diagnostic testing. In N. Frederiksen et al. (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 489-497). Hillsdale, NJ: LEA.
 8. Lohman, D. F., & Ippel, M. J. (1993). Cognitive diagnosis: From statistically based assessment toward theory-based assessment. In N. Frederiksen et al. (Eds.), *Test theory for a new generation of tests* (pp. 41-71). Hillsdale, NJ: LEA.
 9. Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: LEA.
 10. Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.
 11. Marshall, S. P. (1990). Generating good items for diagnostic tests. In N. Frederiksen et al. (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 433-452). Hillsdale, NJ: LEA.
 12. McDonald, R. P. (1989). Future directions for item response theory. *International Journal of Educational Research*, 13, 205-220.
 13. Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen et al. (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: LEA.
 14. Rost, J. (1985). A latent class model for rating data. *Psychometrika*, 50, 37-49.
 15. Rost, J. (1988). Rating scale analysis with latent class models. *Psychometrika*, 53, 327-348.
 16. Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111-121.
 17. Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen et al. (Eds.), *Test theory for a new generation of tests* (PP. 1-17). Hillsdale, NJ: LEA.
 18. Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 104, 385-395.
 19. Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
 20. Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical Pattern classification and recognition approach. *Behoviormetrika*, 19, 73-86.
 21. Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen et al. (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: LEA.
 22. Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and

- pattern classification. *Psychometrika*, 52, 193-206.
23. Tatsuoka, K. K., & Tatsuoka, M. M. (1988). Rule space. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, 8, (pp. 217-220). New York: John Wiley & Sons.
24. Wainer, H. et al. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: LEA.

本文转载自研习信息 11 卷（3 期），7-11 页