心理与行为研究 2015，13(6)：832~838
Studies of Psychology and Behavior

等级反应题组模型在《基于实际操作的老年人日常问题解决能力测验》中的应用 *

陈 勃 邓稳根 李慧琦
（赣南师范学院教育科学学院，赣州 341000）

摘 要 为考查题组模型在实际应用中的性能及对《基于实际操作的老年人日常问题解决能力测验》的项目质量进行分析，选取了 450 个 60 到 84 岁的老年人的测量数据，进行了模型-数据拟合检验，比较了不同模型的拟合优度，并采用等级反应题组模型（GRTM）对测验项目的参数和题组效应进行了估计。结果显示，包含 10 个特殊因子和一个公共因子的双因子模型相对更拟合测量数据，将项目组成 10 个题组后，题组的效应非常显著的题组数达到了 9 个，GRTM 相对于等级反应模型对测验的项目质量作了更好的分析，说明 GRTM 在测量实践中是可行的，未来测验的编制要加强测量模型的研究。

关键词 等级反应题组模型，等级反应模型，局部相依性，老年人，日常问题解决。
分类号 B841.7

1 引言

日常问题解决能力是近年来老年人心理研究中比较活跃的领域之一。日常问题解决涉及到生活中的多个方面，较为一致的看法是把日常问题解决分为完成活动任务和处理人际关系两种类型，也即工具性问题解决和人际性问题解决(Camp, Doherty, Moody-Thomas, & Denny, 1989; Marsiske & Willis, 1995; Blanchard-Fields, Chen, & Norris, 1997)。工具性问题解决（*instrumental problem solving*），指的是个体从事对其生存具有工具性意义的问题解决活动，例如，经典的工具性日常问题解决测验（*Instrumental Activities of Daily Living*, IADL; Lawton, Brody, & Médecin, 1988）就由服药、备餐、理财、打电话、洗衣、家务和出行等七项反映日常问题解决能力的任务组成。人际性问题解决（*interpersonal problem solving*）主要涉及人际情绪情感因素，指处理在社会关系情境中产生的情绪情感反应的问题，如取悦妻子、消除夫妻间矛盾等。人际性问题解决由于涉及情绪情感等非理性因素，具有高度的复杂性及不确定性，因此，难以进行定量评定。目前，对于老年人日常问题解决能力的评估主要涉及的是工具性问题。

而对于老年人工具性问题解决能力的评估方法，大致可以分为两类：一类是基于非实际操作的评估方法，包括自我报告法、代理报告法、纸笔测验法；另一类是基于实际操作的方法。Willis 和 Marsiske（1993）编制的日常问题解决测验（*Everyday Problem-solving Test*, EPT）主要以纸笔测验和附加自评式项目的形式来评估老年人日常问题解决能力；我国研究者（李茵，王大华，申继亮，1997）也将自我报告法与纸笔测验结合起来，编制了适合我国老年人的日常问题解决测验。自我报告和纸笔测验等基于非实际操作的评估方法更多的是反映被试的认知能力，而难以反映老年人在面对实际日常问题时的解决能力，越来越多的研究者主张根据被试的实际操作行为对其日常问题解决能力进行评估。例如，Diehl，Willis 和 Schaie（1995）在 EPT 基础上发展起来的可观察的日常生活任务（*Observed Tasks of Daily Living*, OTDL）测验就是给被试呈现一些生活中的常见物（如蛋糕搅拌器、药品、电话本）作为刺激，要求被试通过现场操作完成 9 项备餐方面的任务，13 项服药方面的任务和 9 项打电话方面的任务，根据被试在这些任务上的表现差异对他们的问题解决能力进行评估。Diehl 等人（2005）对该测验进行了修订，修订版（OTDL-R）将备餐一项改为理财，测验项目数也由原来的 31 项调整为 9 项，新修订的测验测试时

收稿日期：2015-10-23
* 基金项目：国家自然科学基金项目（81160140）。
通讯作者：邓稳根，E-mail：dwengen@163.com。

间更短，任务难度分布范围更大，效度也符合心理学测量学的要求。与基于非实际操作的评估方法相比，这种基于实际操作的评估方法的优点在于与老年人的生活实际及真实的问题解决任务更为接近，不仅保证了较高的内部效度，而且保证了拥有较高的外在效度或生态效度（Allaire, Gamaldo, Ayotte, Sims, & Whitfield, 2009）。

无论是自我报告法、代理报告法、纸笔测验法，还是基于实际操作的评估方法，当前老年人日常问题解决能力的测量都是基于经典测验理论（*Classical Test Theory*, CTT）而进行的。CTT 虽然历史悠久，应用广泛，但 CTT 具有很大的局限性。例如，测验统计量（例如，信度和效度）的值受到被试样本分布的影响，由于老年人的个体差异极大，在测试中很难抽取到有代表性的老年人样本，因此测验统计量的计算结果经常不一致，这严重影响到结果的可靠性和准确性，从而限制了老年人日常问题解决能力的评估结果在实践中的进一步应用。项目反应理论（*Item Response Theory*, IRT）是针对 CTT 的局限而提出来的，相对于 CTT，IRT 具有许多优点。例如，项目参数的估计更少受到被试样本分布的影响，被试能力参数的估计更少受到测验的项目抽样影响，能够更精确地反映测验和项目在不同能力水平被试上的测量误差等等。

尽管 IRT 具有如此多的优点，但它并没有成为当前老年人日常问题解决能力测验编制的理论依据，原因是多方面的，其中一个原因在于当前的大多数 IRT 模型并不适合老年人日常问题解决能力的测量。老年人日常问题解决能力测验往往是由情境依赖性题组（*context-dependent item set*）构成。所谓情境依赖性题组，指的是一组共用一个刺激或共同信息来源的相关试题（Haladyna, 1992; Lee, 2000; Allen & Sudweeks, 2001）。因此，一个情境依赖的题组必然包含一个刺激（*stimulus*）或题干，和一组需以该刺激或题干作为答题依据的试题。IRT 模型的基本假设之一——局部独立性（*local independence*, LI）要求在给定能力水平的情况下，被试正确作答各试题的概率是相互独立的。如果被试在一个试题上的反应，受到能力值以外的因素影响，这时就会产生局部试题相依性（*local item dependence*, LID）。在题组题型中，由于同一个题组内的试题是使用相同一段文字刺激，所以试题间可能相互有关联，被试在作答某一试题时会受到其它试题的作答影响。

根据许多研究者的分析与探讨，忽略试题的局部相依性，将可能造成两种负面结果。第一种是测验信息量的高估或测验标准误的低估（Sireci, Wainer, & Thissen, 1991; Wainer & Thissen, 1996; Yen, 1993）。Yen（1993）指出，假如测验存在局部试题相依性，将会造成几乎 100%信息函数量高估，因而会低估想要达到测验精度所需要的试题数。第二种是参数估计的偏差，Yen（1993）认为如果存在局部试题相依性，平均会有 60%的试题参数估计会有偏差。

有鉴于此，Bradlow，Wainer 和 Wang（1999）通过引入一个新的概念——题组随机效应参数，将标准的两参数逻辑斯蒂克模型（*two Parameter Logistic Model*, 2PLM; Birenbaum, 1968）推广到带有题组效应参数的两参数逻辑斯蒂克题组反应（*Testlet Response*）模型，标志着题组反应理论（*Testlet Response Theory*, TRT）的出现。

TRT 是从 IRT 拓展而来的，TRT 的模型与 IRT 模型最基本的区别在于引入了题组参数。Bradlow，Wainer 和 Wang（1999）将 IRT 模型中线性测量的标准形式

$t_{ij}=\alpha_j(\theta_i-b_j)$

扩展为 TRT 模型中的形式

$t_{ij}=\alpha_j(\theta_i-b_j+\gamma_{id(j)})$

其中，α_j 和 b_j 分别是试题 j 的区分度和难度参数，θ_i 为被试 i 的能力参数，$\gamma_{id(j)}$ 为题组 d_j 中试题 j 和被试 i 的交互作用，即题组的随机效应参数。当 $\gamma_{id(j)}=0$ 时，所有试题均局部独立，反之则存在局部相依。因此，参数 $\gamma_{id(j)}$ 的设定体现了 TRT 模型和与 IRT 模型之间的差别。

Wang，Bradlow 和 Wainer（2002）采用类似的方法将等级反应模型（*Graded Response Model*, GRM; Samejima, 1969）拓展为等级反应题组模型（*graded-response testlet model*, GRTM），并随后开发了相应的软件 SCORIGHT 3.0（Wang, Bradlow, & Wainer, 2004）在一个完全贝叶斯框架内使用马尔科夫链蒙特卡罗（*Markov chain Monte Carlo*, MCMC）方法估计题组模型的参数。

本研究成员根据日常工具性问题的内涵，采用访谈的方法，得出了中国老年人常见的日常工具性问题主要有四大类，分别为服药、打电话、电器使用和理财。每一类问题又包含若干亚类，例如，打电话包含了拨打紧急电话、座机和手机三个亚类。根据各亚类问题，开发了一些实物模型，并根据实

物模型设计了若干个需要通过实际操作才能完成的项目组成测验，并将测验命名为《基于实际操作的老年人日常问题解决能力测验》。由于本研究的主要目的是采用题组模型分析该测验的项目质量，因此对于该测验的编制过程这里并不进行阐述。

该测验的一个明显特点是同一亚类问题的项目共用同一个或几个实物模型，因此，非常符合情景依赖性题组的定义，根据前面的分析可以推断，采用 CTT 和 IRT 模型来分析该测验的测量数据可能会得出错误的结论，相反地，采用 TRT 模型对其数据进行分析则比较合适。但非常遗憾的是，目前研究者对 TRT 模型虽然开展了一些模拟研究，但它在实际测验中的应用很少看到，特别是在心理测验中的应用，目前为止尚未见到。模拟研究的条件毕竟与实际有一定的差距。例如，本研究中所使用的测验与一般的测验不同，测验中每一个项目总是和其它几个项目共用一个或一批实物模型，没有只属于某一个项目而不属于任何其它项目的实物模型，因此，理论上测验都是由题组构成，过去的题组模拟研究中模拟的测验往往是由独立项目和题组项目混合而成，较少见到完全由题组构成的测验。

因此，考查 TRT 模型在《基于实际操作的老年人日常问题解决能力测验》中的应用性不仅对于实际应用是非常重要的，而且对于题组理论的发展和题组模型的完善也具有很大价值。

2 对象与方法

2.1 研究对象

本研究施测的对象为具有生活自理能力的老年人，被试年龄分布范围为 60 至 84 周岁。各年龄段的人数比例尽量与江西省人口普查的结果相一致，即年龄越大的老年人所占比例越小。研究者深入社区、老年人活动中心、老年公寓、福利院等多地进行施测，共发放及回收测验 629 份，回收率 100%。剔除 4 份超龄被试（≧85 岁）完成的无效测验及 179 份数据缺失的测验，还剩有效测验 450 份，有效率 71.54%。其中男性 227 人，女性 223 人；60~64 岁的被试有 138 人，65~69 岁有 108 人，70~74 岁 93 人，75~79 岁 81 人，80~84 岁 30 人；小学文化程度以下 14 人，小学 70 人，初中 143 人，高中 53 人，中专 52 人，大专 42 人，大学 63 人，大学以上 13 人。

2.2 测量工具

研究采用《基于实际操作的老年人日常问题解决能力测验》进行施测，该测验共 46 个项目，其中服用药物分测验 13 项、使用电话分测验 13 项、使用电器分测验 10 项和管理财务分测验 10 项。每个分测验又由若干个子测验组成，全测验共有 10 个不同子测验，分别为药品说明书（8 项）、药品处方单（3 项）、外用药（2 项）、紧急电话（3 项）、座机（4 项）、手机（6 项）、电视遥控器（5 项）、洗衣机（5 项）、现金（6 项）、ATM 机（4 项）。每一个子测验包含一个或一组实物模型刺激，例如，子测验一的刺激是氯化钾缓似片、法莫替丁片、双氯芬酸钠缓释胶囊和醋酸地塞米松片四种药物。46 个项目都要求被试在规定时间内尽力按要求完成操作任务。大多数项目的记分可以分为三个等级水平。第一个等级为被试没有正确完成操作任务得 0 分；第二个等级为被试正确解决问题时，主试给予了被试提示，计 1 分；第三个等级为被试正确解决问题时，主试没有给被试提示，计 2 分。只有项目 38、39、41 和 42（属于理财项目）有 4 个记分等级，被试没有正确完成任务，记 0 分，经提示并使用草稿纸完成任务记 1 分，经提示未使用草稿纸或未经提示但使用草稿纸都记 2 分，未经提示未使用草稿纸记 3 分。

2.3 施测过程

本研究采用的测验是基于实际操作的，属于个别测验，因此在对被试进行施测前需要对主试进行培训，并给每位主试配备一名助手。培训者根据主试指导手册对主试进行培训，使每位主试熟悉测验仪器、测试程序、指导语及评分规则，培训力求细致深入，逐题逐项的反复讲解，以使主试充分掌握施测方法。同时，向助手展示说明测验仪器的使用摆放，以及如何配合主试开展施测。在讲解过程中，培训者需要特别列举说明并模拟实验时可能出现的突发状况，教会他们灵活应对和处理此类事件。

多次培训结束后进行测验施测。实物操作施测是二（主试+助手）对一（被试）施测的，施测对场地要求较高，不能受外界干扰，各组间也须互不干扰，因此要求一个比较安静自然的环境。主试和助手需认真耐心，要严格按照主试指导手册规定进行施测。由于被试差异性较大，测验耗时从半小时到一小时不等，施测结束后给予被试一定的报酬。

2.4 分析方法

采用验证性全息项目因子分析方法使用 R 软件中的 mirt 包（Chalmers, 2012）验证 TRT 模型相对于 IRT 模型更拟合测量数据，进而采用 GRTM 在 SCORIGHT 3.0 软件上对测验的项目参数进行估计

并考查题组效应的大小，同时也采用 GRM 在 MULTILOG 7.03 软件（Thissen, Chen, & Bock, 2003）上对项目参数进行估计。最后，考查 GRTM 模型的项目参数与 GRM 下对应的项目参数之间的关系，采用积差相关系数和散点图两种方式对参数之间的关系进行描述。

3　结果与分析

3.1　模型–数据拟合比较

为了验证 TRT 相对于 IRT 更适合分析本测验的测量数据，本研究比较了四种模型的模型–数据拟合程度：单维 GRM 模型、四维 GRM 模型、包含四个特殊因子的双因子模型和包含 10 个特殊因子的双因子模型。前两个模型属于 IRT 模型范畴，而 TRT 模型则是双因子模型的特殊形式（詹沛达，王文中，王立君，2013）。这四个模型的设定有一定的理论基础。单维的 GRM 假定该测验只测量了日常问题解决能力这一种能力，因此，该测验只有一个维度；四维 GRM 模型则认为这个测验包括四种情景，测量的是四种情景性问题解决能力，因此，测验包含四个维度；双因子模型则将上述两种看法进行了调合，第一个双因子模型认为这个测验实际上测量了一个一般的日常问题解决能力和四个特殊情景的问题解决能力，而第二个双因子模型则认为这个测验实际上测量了日常问题解决能力和 10 个题组构成的方法效应。

本研究采用 AIC 和 BIC 等拟合指数用于比较四个竞争性模型的模型–数据拟合程度。一般而言，AIC 和 BIC 越小，表明该模型相对于其它的模型更拟合测量数据。本研究中四个模型的拟合检验结果如表 1 所示。从表 1 中可以看出，包含一个公共因子和十个特殊因子的双因子模型（即题组模型）相对于其它三个模型对测量数据拟合最好。因此，从验证性全息项目因素分析结果来看，采用题组模型对测量数据进行分析更合理。

表 1　四个模型的拟合结果

	AIC	AICc	BIC	SABIC	Log–likelihood
单维 IRT 模型	25786.08	25918.36	26369.59	25918.94	–12751.04
四维 IRT 模型	26153.14	26285.43	26736.66	26286	–12934.57
一个公因子，四个特殊因子的双因子模型	25294.16	25566.44	26066.7	25470.06	–12459.08
一个公因子，十个特殊因子的双因子模型	25181.33	25453.6	25953.87	25357.23	–12402.66

3.2　题组效应分析

将项目分成 10 个题组后，采用 GRTM 对这 10 个题组的题组效应进行分析。题组效应可以作为题组内项目局部相依程度大小的指标，若不为 0，则可认为题组内题目存在相互干扰（即存在相依性）。但由于题组效应是随机变量，不能直接得到，所以通常采用题组效应的方差作为题组效应大小的指标。Wang，Bradlow 和 Wainer（2002）指出，题组效果的方差若在 0.5 以下可视为具有微量至适度的相依量，大于 0.5 至 1 可认为题组项目局部相依程度较大，大于 1 则意味着完全违背项目间局部独立性（即项目间相依性非常大）。本研究所得的题组效应方差如表 2 所示。从表 2 中可以看出，除了题组 5 的题组效应方差小于 0.50 之外，其余 9 个题组效应方差均高于 0.50，而且所有题组效应方差与其标准误的比值都明显高于 2。表明这些题组的效应非常明显，题组内项目存在相依性，因此有必要将这些效应从参数估计中分离出来。

表 2　10 个题组的题组效应方差及标准误估计值

题组效应	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
方差	0.58	1.35	0.89	0.86	0.38	0.52	1.03	0.71	0.55	0.83
标准误	0.14	0.34	0.34	0.27	0.14	0.14	0.25	0.16	0.15	0.19

3.3　项目质量分析结果

综合上述分析结果，本研究将项目分成10 个题组，GRTM 的项目参数估计结果，如表 3（a）和表 3（b）所示。从表 3 中可以看出，项目的区

分度参数估计值（a）的分布范围在 0.47 到 1.77 之间，只有项目 7、20、40、43 的区分度小于 0.5。表 3 显示测验难度分布范围在–3.92 到 1.82 之间，表明测验更有利于低水平被试的作答。第一个难度水平的估计值（b1）除了第 20 题为正之外，其余均为负，表明绝大多数被试能够在有提示的情形下正确完成操作任务。第二个难度水平的估计值（b2）全部为正，且第 44、45、46 题难度超过 1，这表明第二个难度等级为较难和难，总体来说没有提示就能正确完成操作任务的被试比较少，尤其是第 44、45、46 题正确完成操作任务的被试基本上都是给予了提示。第 38、39、41、42 题第三个难度水平的估计值（b3）为正，难度等级为较难。

表 3（a） 《基于实物操作的老年人日常问题解决能力测验》的项目参数估计值

题号	a	*SE*	b1	*SE*	b2	*SE*	题号	a	*SE*	b1	*SE*	b2	*SE*
1	0.69	0.08	–2.03	0.23	0.98	0.09	22	0.85	0.12	–1.26	0.16	0.16	0.04
2	0.64	0.08	–1.50	0.20	0.68	0.07	23	0.67	0.10	–2.10	0.25	0.20	0.05
3	0.64	0.08	–1.10	0.17	0.43	0.05	24	0.50	0.07	–1.88	0.28	0.09	0.03
4	0.91	0.12	–2.24	0.23	0.68	0.09	25	0.53	0.08	–1.99	0.28	0.15	0.04
5	0.53	0.06	–1.91	0.25	0.97	0.08	26	0.71	0.10	–1.32	0.18	0.08	0.03
6	0.77	0.10	–2.89	0.32	0.66	0.10	27	0.51	0.07	–1.88	0.27	0.32	0.05
7	0.48	0.07	–1.03	0.20	0.18	0.04	28	1.48	0.26	–1.70	0.19	0.65	0.13
8	0.71	0.10	–2.51	0.29	0.35	0.07	29	1.77	0.35	–2.16	0.21	0.92	0.23
9	1.51	0.31	–2.36	0.26	0.84	0.19	30	1.50	0.29	–1.98	0.19	0.25	0.09
10	0.98	0.16	–2.23	0.25	0.40	0.08	31	1.03	0.17	–2.69	0.28	0.19	0.08
11	1.21	0.22	–2.18	0.24	0.36	0.09	32	0.61	0.09	–3.53	0.41	0.65	0.10
12	1.19	0.28	–2.91	0.30	1.02	0.27	33	0.67	0.09	–1.08	0.16	0.65	0.07
13	0.70	0.11	–2.78	0.32	0.26	0.06	34	1.17	0.16	–0.20	0.09	0.48	0.07
14	0.71	0.11	–0.87	0.16	0.18	0.04	35	0.93	0.12	–1.16	0.14	0.48	0.07
15	0.85	0.14	–1.41	0.18	0.09	0.03	36	0.51	0.07	–1.96	0.27	0.66	0.07
16	0.65	0.10	–1.79	0.25	0.15	0.04	37	0.69	0.12	–3.92	0.49	0.26	0.10
17	0.76	0.11	–1.70	0.20	0.21	0.05	40	0.47	0.07	–3.17	0.43	0.36	0.06
18	0.62	0.09	–2.09	0.26	0.11	0.03	43	0.48	0.06	–1.88	0.27	0.74	0.07
19	0.72	0.10	–1.52	0.19	0.12	0.03	44	0.63	0.07	–2.33	0.24	1.49	0.09
20	0.49	0.08	1.82	0.27	0.04	0.02	45	0.95	0.11	–1.01	0.13	1.31	0.12
21	0.77	0.11	–1.85	0.23	0.32	0.06	46	0.85	0.10	–0.57	0.12	1.35	0.10

表 3（b） 《基于实物操作的老年人日常问题解决能力测验》的项目参数估计值

题号	a	*SE*	b1	*SE*	b2	*SE*	b3	*SE*
38	0.53	0.08	–3.78	0.51	0.14	0.06	0.41	0.09
39	0.64	0.08	–1.40	0.19	0.10	0.03	0.58	0.06
41	0.67	0.09	–1.73	0.21	0.10	0.03	0.72	0.07
42	0.58	0.08	–1.81	0.25	0.05	0.02	0.55	0.07

4 讨论

本研究首先对单维 GRM 模型、四维 GRM 模型、包含四个特殊因子的双因子模型和包含 10 个特殊因子的双因子模型的模型–数据拟合程度进行了比较，结果发现包含 10 个特殊因子的双因子模型的模型–数据拟合程度最好，进一步的题组效应检验发现，10 个题组中有 9 个题组的效应方差高于 0.50，表明等级反应题组模型适合分析《基于实物操作的老年人日常问题解决能力测验》的测量数据。

本研究对《基于实物操作的老年人日常问题解决能力测验》的测量数据采用 GRTM 分析，结果

显示项目第一水平的难度参数估计值除了项目 20 为正之外，其余均为负，而第二水平的难度参数估计值全部为正，表明项目的难度设计基本合理，体现了等级的差异。在区分度方面，除了项目 7、20、40、43 的区分度参数低于 0.5（0.50 可被视为区分度良好）之外，其余测验的项目区分度均高于这个值。一般而言，对于常模参照性测验，项目区分度的高低是筛选项目的重要依据，结合区分度和难度值以及项目的关联性，本研究认为项目 20 可以删除，因为该项目难度太大，区分度偏低，与其它项目在内容上不存在依存性，删除该项目不致影响其它项目。项目 40 的难度较低，区分度也不高，因而，也可以考虑删除。第 7 题和 43 题虽然区分度低，但这两个项目的内容都是老年人生活中会遇到的日常问题，同时它们又和其它项目在内容上紧密关联，第 7 题是第 6 题的后续操作，第 43 题是第 44 的前提操作，因此，删除它们会影响到其它项目的作答，可以考虑进行修改。

根据上面的分析，本研究认为未来测验的编制要依赖更合理的测量模型，才能对测验的质量作出更合理的推断，提高测验的科学性水平。

5　结论

TRT 模型较 IRT 模型更适用于《基于实际操作的老年人日常问题解决能力测验》的数据分析，《基于实际操作的老年人日常问题解决能力测验》的项目性能总体上表现良好。未来需要进一步加强题组模型在实际测验中的研究。

参　考　文　献

李茵，王大华，申继亮．(1997)．成人晚期自我觉知的日常问题解决能力及其与日常认知的关系．*心理科学，21*，437–443.

詹沛达，王文中，王立君．(2013)．项目反应理论新进展之题组反应理论．*心理科学进展，21* (12)，2265–2280.

Allen, S., & Sudweeks, R. R. (2001, April) . *Identifying and managing local item dependence in context–dependent item sets*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Allaire, J. C., Gamaldo, A., Ayotte, B. J., Sims, R., & Whitfield, K. (2009) . Mild cognitive impairment and objective instrumental everyday functioning: The everyday cognition battery memory test. *Journal of the American Geriatrics Society, 57* (1) , 120–125.

Birenbaum, A. (1968) . Statistical theories of mental test scores. Blanchard–Fields, F., Chen, Y., & Norris, L. (1997) . Everyday problem solving across the adult life span: influence of domain specificity and cognitive appraisal. *Psychology and aging, 12* (4) , 684.

Blanchard–Fields, F., Chen, Y., & Norris, L. (1997) . Everyday problem solving across the adult life span: Influence of domain specificity and cognitive appraisal. *Psychology and aging, 12* (4) , 684–693.

Bradlow, E. T., Wainer, H., & Wang, X. H. (1999) . A bayesian random effects model for testlets. *Psychometrika, 64* (2) , 153–168.

Camp, C. J., Doherty, K., Moody–Thomas, S., & Denny, N. W. (1989) . Practical problem solving in adults: A comparison of problem types and scoring methods. In J. Sinnott (Ed.) , *Everyday problem solving: Theory and application*. New York: Praeger.

Chalmers, R. P. (2012) . mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48* (6) , 1–29.

Diehl, M., Marsiske, M., Horgas, A. L., Rosenberg, A., Saczynski, J. S., & Willis, S. L. (2005) . The revised observed tasks of daily living: A performance–based assessment of everyday problem solving in older adults. *Journal of Applied Gerontology, 24* (3) , 211–230.

Diehl, M., Willis, S. L., & Schaie, K. W. (1995) . The observed tasks of daily living (OTDL) test. *Psychology and Aging, 10*, 478–491.

Haladyna, T. M. (1992) . Context–dependent item sets. *Educational Measurement: Issues and Practice, 11* (1) , 21–25.

Lawton, M. P., Brody, E. M., & Médecin, U. (1988) . *Instrumental Activities of daily living (IADL)* .

Lee, G. (2000) . A comparison of methods of estimating conditional standard errors of measurement for testlet–based test scores using simulation techniques. *Journal of Educational Measurement, 36* (2) , 91–112.

Marsiske, M., Willis, S. (1995) . Dimensionality of everyday problem solving in older adults. *Psychology and Aging, 10*, 269–280.

Samejima, F. (1969) . *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph No. 17 (pp. 1–100) . Richmond, VA: Psychometric Society.

Sireci, S. G., Wainer, H., Thissen, D. (1991) . On the reliability of testlet–based tests. *Journal of Educational Measurement, 28*, 237–247.

Thissen, D., Chen, W. H., & Bock, R. D. (2003) . *Multilog (version 7) [Computer software]* . Lincolnwood, IL: Scientific Software International.

Wainer, H., & Thissen, D. (1996) . How is reliability related to the quality of testscores? What is the effect of local dependence on reliability?. *Educational Measurement: Issues and Practice, 15* (1) , 22–29.

Wang, X., Bradlow, E. T., & Wainer, H. (2002) . A General Bayesian Model for Testlets: Theory and Applications. *Applied Psychological Measurement, 26* (1) , 109–128.

Wang, X., Bradlow, E. T., & Wainer, H. (2004) . User's Guide for SCORIGHT (Version 3.0) : A computer program for scoring tests built of testlets including a module for covariate analysis. *ETS Research Report Series, 2004* (2) , 1–59.

Willis, S., & Marsiske, M. (1993) . *Manual for the everyday problem test*. University Park: Pennsylvania State University.

Yen, W. (1993) . Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

The Application of the Graded Response Testlet Model into the Performance–based Assessment on Everyday Problem Solving in Chinese Older Adults

Chen Bo, Deng Wengen, Li Huiqi

(School of Education Science, Gannan Normal University, Ganzhou 341000)

Abstract

In order to investigate whether a testlet model can be applied into the performance–based assessment on everyday problem solving in Chinese older adults, 450 older adult participants from 60 to 84 years old from Jiangxi province in China were sampled, including 227 males and 223 females. The goodness–of–fit indexes were tested among the graded response model (GRM) , the multidimensional GRM, and two bifactor models which have four and ten specific factors, respectively. The graded response testlet model (GRTM) and GRM were used to analyze the measurement data. The results showed that the bifactor model with ten specific factors was better fitted to the measurement data than other models, and that when items were incorporated into 10 testlets, there were 9 testlets whose effect size variances were beyond 0.50. It was concluded that GRTM can be applied to analyze the measurement data of the performance–based assessment on everyday problem solving in Chinese older adults.

Key words graded response testlet model, graded response model, local dependence, older adults, everyday problem solving.