This is the pre-published version.

心理科学领域内的客观测量

——Rasch 模型之特点及发展趋势*

晏子

（特殊学习需要及融合教育中心，香港教育学院，香港）

摘　要： Rasch 模型是在国外学术界受到广泛关注和深入研究的一个潜在特质模型。该模型为解决心理科学领域内测量的客观性问题提供了一个可行性很高的解决方案。而国内关于 Rasch 模型的理论探讨和应用研究却并不多见。不同于普通项目反应理论，Rasch 模型要求所收集的数据必须符合模型的先验要求，而不是使用不同的参数来以适应数据的特点。Rasch 模型的主要特点（包括个体与题目共用标尺、线性数据、参数分离）确保了客观测量的实现。未来关于 Rasch 模型的研究方向包括多维度 Rasch 模型、测验的等值与链接、计算机自适应性考试，大型应用测量系统（比如 Lexile 系统）等等。

关键词： Rasch 模型；潜在特质模型；客观测量

分类号： **C8**

Rasch 模型（Rasch, 1960）是由丹麦数学家和统计学家 Georg Rasch（1901-1980）提出的一个潜在特质模型。这一模型以自然科学领域内的客观测量为标杆，为社会科学领域内的测量建立起一套客观标准，以确保测量所提供的信息更为客观和可靠（Bond & Fox, 2007）。经过半个世纪的发展，Rasch 模型已在心理科学领域得到了广泛应用（例如，Merrell & Tymms 2005; Mok, Cheong, Moore, & Kennedy, 2006; Waugh, 2002, 2003; Weaver, 2005）。在国内，虽然早在 80 年代就已经有了关于 Rasch 模型的介绍和研究，但很长一段时间内，这一领域并未赢得学术界足够的重视。笔者作过一个简单的统计，在“中国知网”（1915 至 2008 年）和“中国期刊全文数据库”（1915 至 2009 年）中以”Rasch”为主题进行搜索，总共只找到 93 篇非重复中文文献（搜索日期为 2009 年 11 月 10 日）。文献数量按年份分布如表 1。

表 1 有关 Rasch 模型中文文献的数量分布

年份	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
文献数量（篇）	1	2	9	5	13	63

在 2000 年之后，尤其是最近 5 年，对 Rasch 模型得到了越来越多的重视，研究也日益增多，研究所涵盖的领域包括心理、教育、考试研究、统计、医学、康复等学科。但在已发

* 收稿日期：2009-12-28。 通讯作者：晏子，E-mail: zyan@ied.edu.hk

表的文献中，系统性介绍 Rasch 模型特点以及其发展趋势的仍然很少。少数几篇综述文章多发表于上世纪 90 年代初（例如，Keats, 陈富国，1990; 罗冠中, 1992），无法反映出 Rasch 模型在近二十年的发展。基于此，本文将从基本理论、数学表述、以及主要特点几个方面对 Rasch 模型的进行了讨论，探讨其如何帮助心理科学研究者实现客观测量，并介绍其最新的发展趋势。

1 Rasch 模型的基本理论

作为一种潜在特质模型，Rasch 模型通过个体在题目上的表现（通常表示为原始分数）来测量不可直接观察的、潜在的变量。根据 Rasch 模型原理，特定的个体对特定的题目作出特定反应的概率可以用个体能力与该题目难度的一个简单函数来表示。个体回答某一题目正确与否完全取决于个体能力和题目难度之间的比较。

Rasch 模型是一个理想化的数学模型。它要求所收集的实证数据必须满足事先规定的标准和结构，才能实现客观测量。Wright 和 Stone (1979) 指出，Rasch 模型对于客观测量有两个要求，即：（1）对任何题目，能力高的个体应该比能力低的个体有更大可能作出正确回答；（2）任何个体在容易题目上的表现应该始终好过在困难题目上的表现。

2 Rasch 模型的数学表述

在 Rasch 模型中，个体的能力，题目的难度，以及个体给出正确答案的可能性之间的关系可以由方程（1）来表达。

$$P_{mi}(x_{mi} = 1/\theta_m, \delta_i) = \exp(\theta_m - \delta_i) / [1 + \exp(\theta_m - \delta_i)] \qquad (1)$$

$P_{mi}(x_{mi} = 1/\theta_m, \delta_i)$ 指的是能力为(θ_m) 的个体正确回答$(x = 1)$难度为 (δ_i)的题目的概率。推导过程如下：假如 P_{mi} 是特定个体 m 答对特定题目 i 的概率，很显然，答错的概率便是$(1 - P_{mi})$。同理，特定个体 n 在题目 i 上答对和答错的概率分别是 P_{ni} 和 $(1 - P_{mi})$。如果想比较 m 和 n 的能力，可以用以下比率来表示

$$\frac{P_{mi}(1 - P_{ni})}{(1 - P_{mi})\,P_{ni}}$$

如果测量是客观的，那么，当我们使用题目 j 来代替题目 i 时，上述比例应该保持不变。因为虽然题目不同，但个体 m 和 n 的能力并未没有发生改变，那么两人之间能力的比较也不会发生改变。即

$$\frac{P_{mi}(1 - P_{ni})}{(1 - P_{mi})\,P_{ni}} = \frac{P_{mj}(1 - P_{nj})}{(1 - P_{mj})\,P_{nj}}$$

现在，假设 n 是一个“标准”个体，而题目 j 是一个“标准”的题目。 n 的能力与 j 的难度相等。那么，$P_{nj} = 0.5$。得到

$$\frac{P_{mi}}{1 - P_{mi}} = \frac{P_{mj}\,P_{ni}}{(1 - P_{mj})(1 - P_{ni})}$$

可以定义

$\ln\left(\frac{P_{mj}}{1 - P_{mj}}\right) = \theta_m$ 表示 m 的能力；　　$\ln\left(\frac{1 - P_{ni}}{P_{ni}}\right) = \delta_i$ 表示 i 的难度；

那么

$$\ln\left(\frac{P_{mi}}{1 - P_{mi}}\right) = \theta_m - \delta_i$$

最终得到

$$P_{mi} = \exp(\theta_m - \delta_i) / [1 + \exp(\theta_m - \delta_i)]$$

对于任何特定的 θ_m 和 δ_i，以上方程可以表达为

$$P_{mi}(x_{mi} = 1/\theta_m, \delta_i) = \exp(\theta_m - \delta_i) / [1 + \exp(\theta_m - \delta_i)]$$

3 Rasch 模型与项目反应理论（Item Response Theory, IRT）的关系

IRT 模型是关于个体对题目所作反应的概率与潜在特质之间数学关系的表述。常见的 IRT 模型包括单参数模型、双参数模型、和三参数模型。关于 IRT 模型与 Rasch 模型的关系，存在着两种观点。一种观点由两者在基本理论和数学表述上的相似性出发，认为 Rasch 模型是单参数 IRT 模型的一个特例。单参数 IRT 模型只考虑题目的难度（b）这一个参数，而将 IRT 模型中的其它两个参数－－题目区分度参数（a）和猜测度参数（c）－－设为恒定（a＝1，c＝0）。另一种观点则认为 Rasch 模型有着区别于 IRT 模型的独特的方法论取向。Rasch 模型不将题目区分度和猜测度定义为测量模型中的参数，而将之视为测量过程中应该尽量避免并小心诊断其影响的“噪音”。正如 Wright (1997)所指出，如果同一测验中的题目具有不同区分度，说明存在题目偏见或者该测验具备多维度特性。至于猜测度，它反映的是答题者的一种不可靠性，同样不应该被视为题目的一个参数。正是由于这些参数的影响导致了多参数 IRT 模型在实际应用上的一些缺陷。比如，不同区分度会导致题目之间的相对难度会随着个体能力的变化而改变。在同一测验中，对低能力组来说，A 题难于 B 题，而对于高能力组来说，可能变为 B 题难于 A 题。 这使得多参数 IRT 模型在某种程度上难以达致客观测量的一些要求（比如：参数分离）。

IRT 模型或其他统计方法倾向于使用不同的参数来以适应数据的特点，而 Rasch 模型则要求所收集的数据必须符合模型的先验要求（Andrich, 2004）。这正是 Rasch 模型所强调的“客观测量”的一个关键点。我们可以举一个例子来看一看用参数来适应数据这种方法的不足。有不少研究对体能测验结果进行了因子分析，试图确定体能这一潜在变量的结构（例如, Fleishman, 1964; Marsh, 1993; Ponthieux & Barker, 1963）。而无论是探索性因子分析，还是验证性因子分析，在试图建立客观测量时均有明显缺陷。Marsh (1993) 指出，探索性因子分析使研究人员无法控制最终所得出的因子结构。研究人员无法测试任何先验因子结构，数据所产生的结果便是最终结果。至于验证性因子分析，尽管它可以让研究人员测试其先验因子结构，并提供指标来判断先验因子结构与实证因子结构之间匹配的程度，但也未能达到客观标准。因为数据作为一个“现实”，而因子模型只是用来“解释”这些数据。当模型无法正确地解释数据时，就必须对模型进行修改，对参数进行修订，直到修订后的模型和参数可以很好地解释数据。因此，在上述以数据为本的研究中，要想取得一个稳定的体能因子结构几乎是不可能的，因为各研究中体能测试的样本不同，所使用的体能指标也不同。从这个意义上讲，如果没有建立起一个独立于数据的、客观的尺度，在不同情境所得到的测量结果就不可能进行有意义的比较。有鉴于

此，Rasch 模型设定了客观测量中数据必须满足的先验要求。如果数据不适合 Rasch 模型，首先应该做的是审视数据本身可能存在的问题，而不是改变模型自身参数设置来适应不同的（可能存在问题的）数据。在 Rasch 模型下，不同的研究结果（因子结构、测验量尺、等等）可以直接适用到其他情境下，因此，在不同情境下进行的测量可以在一个稳定和一致的框架内进行解读和沟通。有研究者 (Al-Owidha, 2007）比较了 Rasch 模型和三参数 IRT 模型在同一套学业测验数据上的表现。结果发现，虽然三参数 IRT 模型对数据的拟合度高于 Rasch 模型（这不难理解，因为三参数模型的方法是使用更多参数去使“模型适应数据”，而 Rasch 模型却要求“数据符合模型”），但 Rasch 模型却能提供更稳定、更精确的题目难度参数，以及更好的题目和测验信度。

4 Rasch 模型的主要特点

4.1 个体和题目共用同一把尺

Rasch 模型通过对数转换，将个体和题目在同一单维度尺上进行标定（Wright & Masters, 1982）。基于各自在此单维度连续体上的位置，个体与个体之间、题目与题目之间、个体与题目之间可以方便地进行直接比较。这是 Rasch 模型区别于传统测量方法的一个显著特征，也是实际应用当中最有意义的一个方面。例如：在传统测量方法下，如果 A 题目没有对某学生施测，那么即使该学生回答过类似的另一题目 B，也很难预测其在 A 题目上的表现。然而，Rasch 模型可以解决这一问题。依据各自的能力或难度水平，个体和题目被标定在同一量尺的不同位置上。根据这种相对位置所提供的信息，即使没有真正施测，也可以预测学生在该题目上的表现。

4.2 数据的线性特质

任何观测值都来源于原始数据，但原始数据所提供的却往往并非有效的“量度”，因为从原始数据人们很难作出有价值的推论（Wright, 1997; Wright & Mok, 2000）。Bond 和 Fox (2007)所指出，从原始数据很多时候表示的仅仅是个体或题目的次序，而并非是关于“多少”的问题，也就是说，无法得知不同分数之间的距离，更无法提供分数在比例上的意义，而这恰恰是有效测量的关键所在。心理测验经常使用李科特量表（例如：非常不同意，不同意，同意，非常同意）。学生在此类量表上的原始分数看起来是等距的，但这并不意味着原始分数所代表的心理特质水平也具有等距的意义。因为等距的量度意味着分数每增加一个单位，所代表的特质水平也相应地有一个同等大小的增量。然而事实并非如此。“非常不同意”与“不同意”之间的距离，未必等于“不同意”与“同意”之间的距离。

数据的线性是任何统计方法－－比如因子分析－－的一个基本假设（Wright & Masters, 1982）。然而，很多数据，就象学业考试的原始分数，实质上并不符合线性数据的要求。因此，严格来讲，大部分统计方法并不适用于这种非线性（或非等距）数据。只有将这种数据转换为线性的、等距的数据，才可应用统计方法 (Wright, 1997)。Rasch 模型可以将非线性数据转换成为具有等距意义（对于所测量特质而言）的“logit scale”数据，从而使客观的测量成为可能(Linacre, 2006)。有些学者(例如，Fischer, 1995)甚至认为 Rasch 模型是唯一可行的将次序数据转换为线性数据的方法。

4.3 参数分离

由于个体所得到的原始分数依赖于所施测的题目集，而对分数的解读又依赖于特定施测样本，因此传统测量方法很难用来比较或预测个体在不同测验之间的表现。这是传统测量理论的一个重大缺陷。假设有两份测量同一心理特质的心理测验问卷 A 和 B，一

名学生在 A 卷中得到 80 分，那么他在 B 卷中可以得到多少分？很难预测。即使是同一学生，题目测量的是同一特质，只要题目不同，分数也可能有不同。再举一例：学生甲在 A 卷中得到 80 分，学生乙在 B 卷中也得到 80 分。哪一位学生所对应的心理特质水平更高？很难直接作出判断，因为虽然他们分数相同，但却是在不同测验中得到的，其分数所代表的含义也不同。

为了避免直接对原始分数进行解读所造成的困难，有时会用标准化分数（如 z 分数和 t 分数）代替原始分数来比较在不同测试上的得分。然而，标准分数的计算依赖于所选取的样本。由于不同样本的平均数和标准偏差都不同，意味着基于标准分数的比较只适用于来自同一样本的个体。百分数也有类似的问题。相同的成绩，在不同的常模中所对应的百分数也会不同。

Wright 和 Stone (1979)指出了客观测量两个相辅相成的要求。一个是题目难度的标定必须独立于被试样本的分布，另一个要求是对个体能力的测量必须独立于题目的难度分布。此一特点称为“参数分离”或“参数恒定”（Embretson & Reise, 2000; Wright & Masters, 1982; Wright & Mok, 2000）。 在前文述及之方程（1）中，正确反应的概率只由个体的能力（θ_m）和题目的难度（δ_i）所决定。这意味着 Rasch 模型所提供的个体能力和题目难度参数，是完全独立样本分布或题目难度分布的。因此，Rasch 模型符合客观测量对于参数分离的要求。

然而，需要特别指出的是，在实际应用当中，运用 Rasch 模型对个体能力和题目难度进行标定时，其数值往往会随着题目难度和个体能力的不同组合而改变。这岂不是和“参数分离”的要求不一致吗？其实不然，“参数分离”并非要求每次标定的绝对估值都一样，而是要求个体与题目之间的差异（在潜在特质量尺上的相对位置）保持不变，也就是保持一种相对的恒定。从这个意义上来说，Rasch 测量提供的是关于个体能力和题目难度的等距分数，而不是等比分数。

5 Rasch 模型拟合度

如前所述，Rasch 模型是一个理想的数学模型，在现实的测量中不大可能得到完美的实现。因为再简单的测试，都可能受到无关因素的干扰。例如数学考试，学生的表现除了受数学能力影响之外，还有可能受学生的阅读理解能力（能否读懂题目）的影响。心理测验的成绩主要由所测特质决定，但也可能受施测当时学生的身体状况和意愿，以及其他不可预测的因素影响。虽然测量的复杂性和不完善性是客观存在的，但测量工具开发者和使用者应该知道所收集的数据在何种程度符合测量模型要求。Rasch 分析提供的拟合度指标可以检验实证数据与 Rasch 模型的拟合程度。题目的拟合度指标不好，说明可能存在目标特质之外的其他变量，或者对所测量特质的定义不恰当。

很多运行 Rasch 分析的计算机程序（例如，WINSTEPS，ConQuest）提供两种形式的卡方拟合指标：Outfit Mean Square (Outfit MNSQ) 和 Infit Mean Square (Infit MNSQ)。这些拟合指标都是由残差计算而来。Outfit MNSQ 是残差的均方。Infit MNSQ 则是加权（以方差为加权系数）后的残差均方。Outfit MNSQ 对极端值（异常数据）比较敏感，因为极端值会产生的较大的残差。而 Infit MNSQ 对题目难度与个体能力水平相当的数据较为敏感，因为此类数据方差（加权系数）较大(Smith, 2002)。Outfit MNSQ 和 Infit MNSQ 的取值范围介于 0 到正无穷大。理想值为 1，意味着实际数据完全与 Rasch 模型相拟合。大于 1（underfit）表示实证数据的变异数多于 Rasch 模型的预期；小于 1（overfit）表示实证数据的变异数少于 Rasch 模型的预期。从测量的角度来看， underfit（大于 1）的数据对测量客观性的负面影响要大过 overfit（低于 1）的数据。 Underfit 是由杂乱无章的

答案所造成，会直接损害测量的质量。而 overfit 虽然可能会降低测量的效率，但对测量质量的影响反而不大(Bond & Fox, 2007)。Infit MNSQ 和 Outfit MNSQ 可接受的取值范围在很大程度上取决于研究目的。Linacre (2006)建议取 0.5 至 1.5 的范围，但很多研究选取了更为严格的标准，例如，0.7 至 1.3（Mok et al., 2006; Zhu & Cole, 1996）或 0.8 至 1.4（Wolfe & Chiu, 1999）。Infit 和 Outfit 指标也有标准化的形式，分别表达为 Infit ZSTD 和 Outfit ZSTD。Infit ZSTD 和 Outfit ZSTD 服从 *t* 分布，理想值为 0，标准差为 1。

不过，在 Rasch 分析中对于拟合指标的使用必须谨慎。Wright 和 Panchapakesan (1969)指出，在测验发展过程中，简单地删除拟合指标不好的题目并非值得提倡的做法。测验设计者应该仔细审查这些拟合指标不好的题目，找出可能对其产生影响的其他因素，如区分度和猜测效应的影响。Bond 和 Fox (2007)也建议利用拟合度指标来查找表现异常的题目和个体，而不是将它们作为决定是否删除某个题目的简单标准。Smith (2002)指出，应该把实证数据对测量模型的拟合程度看作是一个连续体，而不是一个简单是或否的问题。换句话说，“拟合”与“不拟合”之间并没有森然的壁垒，应该根据不同情况选择合适的标准。

6 Rasch 模型的发展趋势

如何真正实现测量的客观性一直是困扰心理科学，乃至所有社会科学研究者和实践者的问题。Rasch 模型在解决这个问题上实现了很大的突破，其坚实的理论基础，简单的数学表述也确保了它广泛的应用前景。Rasch 模型在诸多方面与 IRT 模型相类似，但却从根本上避免了多参数 IRT 模型在应用上所固有的缺陷。除了心理科学领域，关于 Rasch 模型的研究和应用还大量出现于教育领域（例如，Ito, Sykes, & Yao, 2008; Liu & Wilson, 2009; Tong & Kolen, 2007），卫生和医学领域（例如，Hsueh, Wang, Sheu, & Hsieh, 2004; Strong, Kahler, Ramsey, & Brown, 2003; Tesio, 2003），体育和运动科学领域（例如，Bowles & Ram, 2006; Hands & Larkin, 2001; Heesch, Masse, & Dunn, 2006；Zhu, 2001; Zhu & Cole, 1996），等等。

Rasch 模型从产生至今已有半个世纪，但仍保有旺盛的生命力，并处于持续不断的发展之中。多维度 Rasch 模型（Multidimensional Rasch Model）是其中一个很重要的趋势。比如运用多维度 Rasch 模型对“国际学生评价项目”（Programme for International Student Assessment，PISA）*数据的分析（例如，Liu & Wilson, 2009）；对包含不同分量表的测验数据进行分析（例如，Cheng, Wang, & Ho, 2009）；等等。这里的多维度并不是对 Rasch 模型单维度要求的一种颠覆，而是一种发展。在多维度 Rasch 模型里，对同一维度的个体能力和题目难度的标定仍然固守单维度原则，但与此同时，它充分利用相关维度特质（或相关分量表）所提供的有用信息，以提高测验的效率和对目标特质测量的精确度。多维度 Rasch 模型在某种程度上解决了单维度模型分析多维度测验数据时遇到的信、效度问题（Rost & Carstensen, 2002; Yao & Schwarz, 2006），也使测验在涵盖较为广阔范围内容的同时，也有较高的测验精确度（Cheng et al., 2009），从而极大地延伸了 Rasch 模型的应用空间和前景。

测验的等值和链接（Test equating and linking）是 Rasch 应用的另一个热点研究领域。测验的等值与链接是指将不同测验中取得的分数转化为可以互相替换或比较的分数的统计过程。等值主要处理内容相同而难度不同的测验，而链接则用来处理内容和难度都不相同的测验(Kolen & Brennan, 2004)。越来越多的研究着眼于运用 Rasch 模型建立一把垂直量尺（vertical scale）（例如， Custer, Omar, & Pomplun, 2006; Hanson & Beguin, 2002; Ito et al., 2008; Pomplun, Omar, & Custer, 2004; Tong & Kolen, 2007）。比如，常识告诉我们小学二年级学生的数学能力应该比一年级学生高，但要想确切知道他们之间的数学能力差距，却很困难。

* 由 OECD（经济合作组织）组织在全球范围进行的一项大型学生学习质量比较研究项目。

因为不同年级的考卷题目所测量的内容和/或题目的难度水平不同，因此所得到的分数无法直接比较。如果构建一把可以测量不同年级水平的数学能力的垂直量尺，将在不同试卷上得到的分数放在同一把量尺上进行比较，就可以知道不同年级学生的数学能力差异，跟踪学生在数学能力上的发展。然而，构建这种垂直量尺的尝试受到许多因素的影响，比如数据收集方案（通用题目设计或逐级共用题目设计）、建尺方法（同时标定或分级标定）、甚至所使用的电脑程序（WINSTEPS、BILOG-MG、或其它程序）。是否存在所谓“最佳方法”，还没有达成一致。

基于 Rasch 模型的计算机自适应性考试（Computer Adaptive Testing, CAT）已成为当今教育测量研究与实践的一个重要发展方向。传统考试方法要求所有考生作答完全一样的题目。背后的一个假设是，任何题目对全体考生提供的评价信息是一样的。而事实并非如此，对某一水平考生有用的题目，对另一水平的考生来说可能完全没有意义。CAT 则根据考生不同的能力水平，提供不同的测验题目，以一种最有效、最经济的方法来标定考生的能力。Rasch 模型在实现 CAT 的各个方面，包括试题库的建设，测验题目难度的标定，题目或测验之间的等值，对“作弊策略”的侦测，以及最后的评分，都扮演着重要角色（例如， Gershon & Bergstrom, 1995; Scalise, 2004; Styles & Andrich, 1993）。

对于 Rasch 模型在实现客观测量中的作用，除了持续不断的理论探讨之外，也越来越多地得到了实际应用的佐证。Lexile 系统（Stenner, Sanford, & Burdick, 2007）便是其中较为成功的一个范例。Lexile 是一个英文阅读评估系统，其基础是基于 Rasch 模型发展而来的针对个体阅读能力和文章阅读难度的 Lexile 量尺。这把量尺有固定的原点和相等的测量单位，可以提供关于个体英文阅读能力和英文阅读材料（包括段落、文章、甚至整本书）的难度水平的客观信息。利用这些信息，可以将个体的阅读能力与阅读材料的难度水平进行匹配，从而更好地促进阅读能力的发展。Lexile 系统现阶段主要还是应用于以英文为母语的群体中，但据笔者所了解的情况，针对中文阅读的 Lexile 系统也正在发展当中。

有批评者认为 Rasch 模型的问题在于太过“完美”，导致在现实世界中的测量很难真正实现。某种程度上来说，这不是 Rasch 模型所独有，而是所有数学模型共有的问题。所谓模型，是排除了所有干扰之后的理想状态，这在本质上就决定了模型在现实世界中不可能百分之百实现。这也是为什么要检验模型与实证数据是否吻合，为什么需要拟合度指标。真正的问题在于，很多数学模型过于复杂，对于实践工作的指导意义不大。Rasch 模型是一个相对简单的模型，以一种最有效率的方式规定了客观测量所需要满足的条件。因此具有极大的实践　指导意义。对于关注 Rasch 模型并有兴趣进行相关研究的同仁来说，如何在进一步推动 Rasch 模型理论发展的同时，将先进的测量技术和结果解读方法介绍给测验的直接施测者和使用者（比如心理测验使用者、一线教师、以及大型考试管理者），以帮助实践工作，应该是今后的重点工作方向。

参考文献：

Keats, J. A., 陈富国. (1990). Rasch 的测验理论. *心理学报, 23*, 267-271.

罗冠中. (1992). Rasch 模型及其发展. *教育研究与实验, 1992 年第 2 期,* 40-43.

Al-Owidha, A. A. (2007). *A comparison of the Rasch model and the three-parameter logistic model applied to the quantitative subtest of the General Aptitude Test, Saudi Arabia* (Unpublished Doctoral dissertation). University of Denver, Colorado, USA.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42,* 1-16.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (*2nd ed.*)*. Mahwah, N.J.: Erlbaum.

Bowles, R. P., & Ram, N. (2006). Using Rasch measurement to investigate volleyball skills and inform coaching. *Journal of Applied Measurement, 7*(1)*,* 39-54.

Cheng, Y. Y., Wang, W. C., & Ho, Y. H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement, 69,* 369-388.

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133–149.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah: Lawrence Erlbaum.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer Verlag.

Fleishman, F. A. (1964). *The structure and measurement of physical fitness.* NJ: Prentice-Hall.

Gershon, R., & Bergstrom, B. (1995). *Does cheating on CAT pay: NOT!* Paper presented at the Annual Meeting of the American Educational Research Association, April 18-22, 1995, San Francisco, CA.

Hands, B., & Larkin, D. (2001). Using the Rasch measurement model to investigate the construct of motor ability in young children. *Journal of Applied Measurement, 2*(2), 101-120.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Heesch, K. C., Masse, L. C., & Dunn, A. L. (2006). Using Rasch modeling to re-evaluate three scales related to physical activity: Enjoyment, perceived benefits and perceived barriers. *Health Education Research, 21*. Retrieved October 3, 2006, from http://her.oxfordjournals.org/cgi/reprint/cyl054v1

Hsueh, I. P., Wang, W. C., Sheu, C. F., & Hsieh, C. L. (2004). Rasch analyses of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke, 35*(3), 721-726.

Ito, K., Sykes, R. C. & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21*(3), 187-206.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs.* Chicago, IL: Winsteps.com.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math Assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*(2), 164-184.

Marsh, H. W. (1993). The multidimensional structure of physical fitness: Invariance over gender and age. *Research Quarterly for Exercise and Sport, 64*(3), 256-273.

Merrell, C., & Tymms, P. (2005). Rasch analyses of inattentive, hyperactive and impulsive behaviour in young children and the link with academic achievement. *Journal of Applied Measurement, 6*(1), 1-18.

Mok, M. M. C., Cheong, C. Y., Moore, P. J., & Kennedy, K. J. (2006). The development and validation of the Self-directed Learning Scales (SLS). *Journal of Applied Measurement, 7*(4), 418-449.

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG–MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64*, 600–616.

Ponthieux, N. A., & Barker, D. G. (1963). An analyses of the AAHPER Youth Fitness Test. *Research Quarterly, 34,* 525-526.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institute.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26,* 42-56.

Scalise, K. M. (2004). *BEAR CAT: Toward a theoretical basis for dynamically driven content in computer-mediated environments* (Unpublished Doctoral dissertation). University of California, Berkeley, California, USA.

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analyses of residuals. *Journal of Applied Measurement, 3*(2), 205-231.

Stenner, A. J., Burdick, H., Sanford, E. E. & Burdick, D. S. (2007). *The Lexile framework for reading: Technical report.* MetaMetrics, Inc.

Strong, D. R., Kahler, C. W., Ramsey, S. E., & Brown, R. A. (2003). Finding order in the DSM-IV nicotine dependence syndrome: A Rasch analyses*. Drug and Alcohol Dependence, 72*(2), 151-162.

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's Progressive Matrices in both the Pencil-and-Paper and Computer-Adaptive-Testing formats. *Educational and Psychological Measurement, 53*(4), 905-925.

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analyses as a tool for rehabilitation research. *Journal of Rehabilitation Medicine, 35*(3), 105-115.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227–253.

Waugh, R. F., (2002). Creating a scale to measure motivation to achieve academically: Linking attitudes and behaviours using Rasch measurement. *British Journal of Educational Psychology, 72*(1)*,* 65-86.

Waugh, R. F. (2003). Measuring attitudes and behaviors to studying and learning for university students: a Rasch measurement model analysis. *Journal of Applied Measurement, 4*(2), 164-180.

Weaver, C. (2005). Using the Rasch model to develop a measure of second language learners' willingness to communicate within a language classroom. *Journal of Applied Measurement, 6*(4), 396-415.

Wolfe, E. W., & Chiu, C. W. T. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3*(4)*,* 360-381.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analyses.* Chicago, IL: MESA Press.

Wright, B. D., & Mok, M. M. C. (2000). Rasch models overview. *Journal of Applied Measurement, 1*(1)*,* 83-106.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analyses. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30,* 469-492.

Zhu, W. (2001). An empirical investigation of Rasch equating of motor function tasks. *Adapted Physical Activity Quarterly, 18*(1)*,* 72-89.

Zhu, W., & Cole, E. L. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research Quarterly for Exercise and Sport, 67*(1), 24-34.

Objective Measurement in Psychological Science:

An Overview of Rasch Model

YAN Zi

Centre for Special Needs and Studies in Inclusive Education,
The Hong Kong Institute of Education

Abstract: Rasch model is a latent trait model which has drawn international interest among researchers. It provides a promising solution to ensure the objective measurement in psychological science. However, the research and applicatoin of Rasch model are not as popular as expected among domestic scholars. Unlike more general IRTs that adopt a "the model fits data" position and use different parameters to accommodate the idiosyncrasies of the data set, the Rasch model requires that "data fit the model". Its unique features including the same metric shared by persons and items, data linearity, and parameter separation ensure the achievement of objective measurement. The foci of future development of Rasch model include multidimensional Rasch model, test equating and linking, computer adaptive testing, and Rasch-based measurement system such as Lexile framework.

Key words: Rasch model, latent trait model, objective measurement